DISCRETE ILL-POSED LEAST-SQUARES PROBLEMS WITH A SOLUTION NORM CONSTRAINT

M. E. HOCHSTENBACH[*], N. MCNINCH[†], AND L. REICHEL[‡]

Dedicated to Heinrich Voss on the occasion of his 65th birthday.

Abstract. Straightforward solution of discrete ill-posed least-squares problems with error-contaminated data does not, in general, give meaningful results, because propagated error destroys the computed solution. Error propagation can be reduced by imposing constraints on the computed solution. A commonly used constraint is the discrepancy principle, which bounds the norm of the computed solution when applied in conjunction with Tikhonov regularization. Another approach, which recently has received considerable attention, is to explicitly impose a constraint on the norm of the computed solution. For instance, the computed solution may be required to have the same Euclidean norm as the unknown solution of the error-free least-squares problem. We compare these approaches and discuss numerical methods for their implementation, among them a new implementation of the Arnoldi–Tikhonov method. Also solution methods which use both the discrepancy principle and a solution norm constraint are considered.

Key words. Ill-posed problem, regularization, solution norm constraint, Arnoldi–Tikhonov, discrepancy principle

AMS subject classifications. 65F10, 65F22, 65R30.

1. Introduction.

Minimization problems with a solution norm constraint,

$$\min_{\|L\boldsymbol{x}\|\leq\Delta} \|A\boldsymbol{x}-\widetilde{\boldsymbol{b}}\|, \quad A\in\mathbb{R}^{m\times n}, \quad L\in\mathbb{R}^{p\times n}, \quad \boldsymbol{x}\in\mathbb{R}^{n}, \quad \widetilde{\boldsymbol{b}}\in\mathbb{R}^{m}, \quad m\geq n\geq p, \qquad (1.1)$$

where $\|\cdot\|$ denotes the Euclidean vector norm, the matrix L is of full row rank, and Δ is a user-specified constant, arise in a variety of applications, including data smoothing [21, 22, 26], approximation by radial basis functions [30], and in ill-posed problems [3, 4, 16, 24, 25]. These references describe several numerical methods; further solution techniques are presented by Gander [7], Golub and von Matt [8], and Lampe, Rojas, Sorensen, and Voss [13].

This paper is concerned with the solution of least-squares problems (1.1) with a matrix A with many singular values of different orders of magnitude close to the origin. This makes the matrix severely ill-conditioned; in particular, A may be singular. Least-squares problems with such a matrix are referred to as discrete ill-posed problems. They arise, for instance, from the discretization of ill-posed problems, such as Fredholm integral equations of the first kind with a smooth kernel. The vector $\widetilde{\boldsymbol{b}}$ represents available measured data, which is assumed to be contaminated by an error $\widetilde{\boldsymbol{e}}\in\mathbb{R}^m$. The latter may stem from measurement and discretization errors. We refer to the vector $\widetilde{\boldsymbol{e}}$ as "noise."

In many applications, the matrix L is the identity matrix I, a discrete approximation of a differential operator, or a projection operator. In the latter cases, the minimization problem (1.1) often can be transformed to standard form, i.e., to an equivalent minimization problem with $L=I$; see, e.g., [6, 17, 20], as well as [10, Section 2.3] for discussions and examples. Therefore many minimization problems (1.1)

[*] Version June 12, 2011. Department of Mathematics and Computer Science, Eindhoven University of Technology, PO Box 513, 5600 MB, The Netherlands, `www.win.tue.nl/∼hochsten`

[†] Department of Mathematical Sciences, Kent State University, Kent, OH 44242, USA. E-mail: `nmcninch@kent.edu`

[‡] Department of Mathematical Sciences, Kent State University, Kent, OH 44242, USA. E-mail: `reichel@math.kent.edu`

of interest can be investigated by studying problems in standard form. We henceforth will assume that the problem (1.1) has been transformed to standard form, i.e., that $L = I$.

It is convenient to introduce the unknown noise-free vector $\boldsymbol{b} \in \mathbb{R}^m$ associated with $\widetilde{\boldsymbol{b}}$, i.e.,

$$\widetilde{\boldsymbol{b}} = \boldsymbol{b} + \widetilde{\boldsymbol{e}}.$$

We would like to determine the minimal-norm solution, $\widehat{\boldsymbol{x}} \in \mathbb{R}^n$, of the unavailable noise-free minimization problem

$$\min_{\|\boldsymbol{x}\| \leq \Delta} \|A\boldsymbol{x} - \boldsymbol{b}\|$$

by computing a suitable approximate solution of the available noise-contaminated least-squares problem (1.1) (with $L = I$). We are particularly interested in the situation considered in [4, 13, 16, 24] when

$$\Delta = \|A^\dagger \boldsymbol{b}\|, \tag{1.2}$$

where $A^\dagger$ denotes the Moore–Penrose pseudoinverse of A. Then $\widehat{\boldsymbol{x}} = A^\dagger \boldsymbol{b}$. Thus, we are interested in the situation when $\|\widehat{\boldsymbol{x}}\|$ is known, but the $\widehat{\boldsymbol{x}}$ is not. More generally, our investigation sheds light on regularization by explicitly bounding the norm of the computed solution.

The minimal-norm solution of the unconstrained noise-contaminated least-squares problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \|A\boldsymbol{x} - \widetilde{\boldsymbol{b}}\| \tag{1.3}$$

can be expressed as

$$\widetilde{\boldsymbol{x}} = A^\dagger \widetilde{\boldsymbol{b}} = A^\dagger(\boldsymbol{b} + \widetilde{\boldsymbol{e}}) = \widehat{\boldsymbol{x}} + A^\dagger \widetilde{\boldsymbol{e}}.$$

Due to the severe ill-conditioning of A, the solution $\widetilde{\boldsymbol{x}}$ is typically dominated by propagated error $A^\dagger \widetilde{\boldsymbol{e}}$ of norm much larger than $\|\widehat{\boldsymbol{x}}\|$. We therefore may assume that $\|\widetilde{\boldsymbol{x}}\| > \Delta$. Thus, we are concerned with the solution of the constrained minimization problem

$$\min_{\|\boldsymbol{x}\| = \Delta} \|A\boldsymbol{x} - \widetilde{\boldsymbol{b}}\|. \tag{1.4}$$

The purpose of the constraint is to reduce the amount of propagated noise in the computed solution. The constrained problem (1.4) is equivalent to the penalized unconstrained minimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \{\|A\boldsymbol{x} - \widetilde{\boldsymbol{b}}\|^2 + \mu\, \|\boldsymbol{x}\|^2\} \tag{1.5}$$

for a suitable Lagrange multiplier $\mu > 0$; see Section 2 for details. This minimization problem also is obtained when applying Tikhonov regularization to the unconstrained problem (1.3). The minimization problem (1.5) has the solution

$$\boldsymbol{x}_\mu = (A^T A + \mu I)^{-1} A^T \widetilde{\boldsymbol{b}}, \tag{1.6}$$

where A^T denotes the transpose of A.

Discrete ill-posed problems are effectively underdetermined. Therefore it can be beneficial to impose known properties of the desired solution $\widehat{\boldsymbol{x}}$ on the computed solution during the solution process. In particular, it may be beneficial to require the computed solution to be of norm (1.2), when the latter quantity is available.

The discrepancy principle furnishes another approach to reduce the propagated error in the computed solution. Assume that a bound ε for the norm of $\widetilde{\boldsymbol{e}}$ is available, i.e.,

$$\|\widetilde{\boldsymbol{e}}\| \leq \varepsilon, \tag{1.7}$$

and that $\boldsymbol{b} \in \mathcal{R}(A)$, where $\mathcal{R}(A)$ denotes the range of A. The discrepancy principle then prescribes that the parameter μ in (1.5) be chosen so that

$$\|A\boldsymbol{x}_\mu - \widetilde{\boldsymbol{b}}\| = \eta\varepsilon, \tag{1.8}$$

where $\eta > 1$ is a user-specified constant independent of ε. With $\mu = \mu(\varepsilon)$ determined by (1.8), one can show that

$$\boldsymbol{x}_\mu \to \widehat{\boldsymbol{x}} \quad \text{as} \quad \varepsilon \searrow 0; \tag{1.9}$$

see, e.g., [9, 12] for proofs in a Hilbert space setting. The constant η is required in these proofs.

We note that the vector (1.6) determined by Tikhonov regularization (1.5), with μ chosen so that (1.8) holds, satisfies

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{x}\| \quad \text{with constraint} \quad \|A\boldsymbol{x} - \widetilde{\boldsymbol{b}}\| = \eta\varepsilon.$$

This can be shown with the aid of Lagrange multipliers. We conclude that Tikhonov regularization may be applied to compute the solution of either (1.4) or (1.8), depending on the choice of the regularization parameter μ.

The equivalence of the least-squares problems (1.4) and (1.5) implies that the discrepancy principle can be implemented by solving (1.4) for a suitable $\Delta = \Delta(\mu)$, where $\mu = \mu(\varepsilon)$; see Section 2 for details. When $\varepsilon > 0$, the discrepancy principle corresponds to a value $\Delta = \Delta(\mu(\varepsilon))$ that is smaller than (1.2). Nevertheless, numerical examples of Section 5 show the constraint (1.2) often to give about as accurate approximations of $\widehat{\boldsymbol{x}}$ as the discrepancy principle (1.8).

This paper has several aims. Section 2 discusses properties of the minimization problem (1.4) and describes solution methods based on the standard and range-restricted Arnoldi processes. In particular, the section considers an application of the range-restricted Arnoldi decomposition method described in [18] to Tikhonov regularization. This decomposition requires the computed approximate solution of (1.4) to live in $\mathcal{R}(A)$. The Arnoldi–Tikhonov method so obtained improves on the scheme described in [15]. Section 3 discusses how both the constraint $\|\boldsymbol{x}\| = \Delta$, with Δ given by (1.2), and the constraint (1.8) can be applied simultaneously. A sensitivity analysis is provided in Section 4, and numerical examples are presented in Section 5. We compare a Tikhonov regularization method based on the standard Arnoldi process and a scheme based on the LSTRS method recently described by Lampe et al. [13] for problems (1.4) with $m = n$ and an error-free vector $\widetilde{\boldsymbol{b}}$, i.e., $\widetilde{\boldsymbol{b}} = \boldsymbol{b}$. The LSTRS-based method uses the nonlinear Arnoldi process presented by Voss [29]. We also compare with a scheme by Li and Ye [16]. None of the iterative methods in our comparison require the evaluation of matrix-vector products with A^T. This feature is important for problems for which it is difficult to evaluate these matrix-vector products. For instance, in large-scale nonlinear minimization problems when A is the Jacobian matrix, the evaluation of matrix-vector products with A may be much cheaper than the evaluation of matrix-vector products with A^T; see, e.g., [5]. We are interested in iterative methods that are based on the Arnoldi process, because they can be applied when A^T is not available, and they may require fewer matrix-vector product evaluations than iterative methods that require matrix-vector products with both the matrices A and A^T; see, e.g., [2, 15] for illustrations. The requirement of iterative methods based on the Arnoldi process that the matrix A be square can be circumvented by zero-padding. This is practical at least when m and n in (1.1) are of about the same size. The computed examples of Section 5 illustrate the benefit of using range-restricted Arnoldi methods when $\widetilde{\boldsymbol{b}}$ contains a nonnegligible amount of noise. Concluding remarks can be found in Section 6.

This paper extends the approach advocated by Lampe et al. [13] for the solution of ill-posed problems (1.4) with a square matrix A in several ways: i) the vector $\widetilde{\boldsymbol{b}}$ is allowed to be contaminated by noise, ii) a range-restricted Arnoldi decomposition is applied, and iii) the constraints in (1.4) and (1.8) are applied simultaneously. Numerical experiments illustrate that the constraint that the computed solution be of norm $\|\widehat{\boldsymbol{x}}\|$ may yield a better approximation of $\widehat{\boldsymbol{x}}$ than the discrepancy principle. This observation is believed to be new. The analysis of Section 4 shows how sensitive the computed solution is to the value of Δ in (1.4).

We conclude this section with some comments on alternative solution methods for (1.4). When the least-squares problems (1.4) or (1.5) are of small to moderate size, they can be solved conveniently by the use of the singular value decomposition (SVD) of A. Large-scale problems can be solved by application of a few steps of Lanczos bidiagonalization; see, e.g., [4, 8]. The latter approach requires evaluation of matrix-vector products with both the matrices A and A^T. An application of the LSTRS method, which does not use the nonlinear Arnoldi process, is described in [24].

This paper blends linear algebra and ill-posed problems, areas in which Heinrich Voss over the years has made numerous important contributions; see, e.g., [13, 14, 28, 29]. It is a pleasure to dedicate this paper to him.

2. Solution norm constraint. We first establish the connection between the constrained minimization problem (1.4) and the penalized unconstrained minimization problem (1.5). This connection implies that methods developed for Tikhonov regularization of linear discrete ill-posed problems can be adapted to solve (1.4). The following result can be shown with the aid of Lagrange multipliers.

Proposition 2.1. *Assume that $0 < \Delta < \|A^\dagger \widetilde{\boldsymbol{b}}\|$. Then the constrained minimization problem (1.4) has a unique solution $\boldsymbol{x}_{\mu_\Delta}$ of the form (1.6) with $\mu_\Delta > 0$.*

We turn to the dependence of $\|\boldsymbol{x}_\mu\|$ on μ. It is convenient to introduce the function

$$\psi(\mu) = \|\boldsymbol{x}_\mu\|^2, \qquad \mu > 0. \tag{2.1}$$

Proposition 2.2. *The function (2.1) can be written as*

$$\psi(\mu) = \widetilde{\boldsymbol{b}}^T A(A^T A + \mu I)^{-2} A^T \widetilde{\boldsymbol{b}}, \tag{2.2}$$

Let $A^T \widetilde{\boldsymbol{b}} \neq \mathbf{0}$. Then $\psi(\mu)$ is strictly decreasing and convex for $\mu > 0$. Moreover, the *equation*

$$\psi(\mu) = \tau \tag{2.3}$$

has a unique solution $0 < \mu < \infty$ *for any* $0 < \tau < \|A^\dagger \widetilde{\boldsymbol{b}}\|^2$.

Proof. Substituting (1.6) into (2.1) yields (2.2). The stated properties of $\psi(\mu)$ and of equation (2.3) can be shown by substituting the singular value decomposition of A into (2.2). □

We also are interested in the function

$$\phi(\mu) = \|\widetilde{\boldsymbol{b}} - A\boldsymbol{x}_\mu\|^2, \qquad \mu > 0. \tag{2.4}$$

Proposition 2.3. *The function (2.4) allows the representation*

$$\phi(\mu) = \widetilde{\boldsymbol{b}}^T (\mu^{-1} AA^T + I)^{-2}\, \widetilde{\boldsymbol{b}}. \tag{2.5}$$

Assume that $A^T\widetilde{\boldsymbol{b}} \neq \boldsymbol{0}$. *Then* $\phi(\mu)$ *is strictly increasing for* $\mu > 0$, *and the equation*

$$\phi(\mu) = \tau \tag{2.6}$$

has a unique solution $0 < \mu < \infty$ *for* $\|P_{\mathcal{N}(A^T)}\widetilde{\boldsymbol{b}}\|^2 < \tau < \|\widetilde{\boldsymbol{b}}\|^2$, *where* $P_{\mathcal{N}(A^T)}$ *denotes the orthogonal projector onto* $\mathcal{N}(A^T)$, *the null space of* A^T. *In particular, if* A *is of full rank, then* $\|P_{\mathcal{N}(A^T)}\widetilde{\boldsymbol{b}}\| = 0$.

Proof. Substituting (1.6) into (2.4) and using the identity

$$I - A(A^TA + \mu I)^{-1}A^T = (\mu^{-1}AA^T + I)^{-1}, \qquad \mu > 0,$$

shows (2.5). The properties of equation (2.6) follow by substituting the singular value decomposition of A into (2.5). □

Proposition 2.3 shows that when ε is increased in (1.8), the corresponding value of μ, such that $\boldsymbol{x}_\mu$ satisfies (1.8) also increases. By Proposition 2.2 the norm $\|\boldsymbol{x}_\mu\|$ then decreases. Indeed, for any $\eta\varepsilon > 0$, the solution $\boldsymbol{x}_\mu$ of (1.8) satisfies $\|\boldsymbol{x}_\mu\| < \|\hat{\boldsymbol{x}}\|$; see, e.g., [12, Section 2.5] for a proof in Hilbert space. In particular, the solution of (1.4) with Δ defined by (1.2) is of larger norm than the solution $\boldsymbol{x}_\mu$ of (1.8) for any $\eta\varepsilon > 0$.

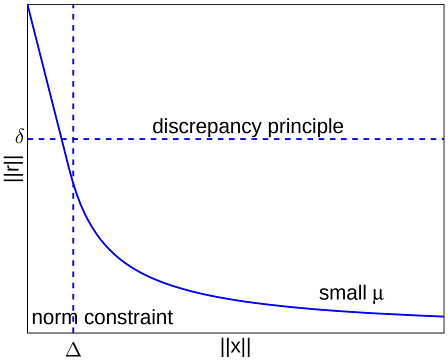


Fig. 2.1. *Example 2.1: The relation between* δ, Δ, μ, $\|\boldsymbol{x}_{\mu(\delta)}\|$, $\|\boldsymbol{x}_{\mu(\Delta)}\|$, *and the residual error.*

Example 2.1. Let Δ be given by (1.2) and $\mu = \mu(\Delta)$ be such that $\|\boldsymbol{x}_{\mu(\Delta)}\| = \Delta$. Similarly, let $\mu = \mu(\delta)$ be determined by (1.8) with $\delta = \eta\varepsilon$. Then for $\delta > 0$, we have

$$\mu(\Delta) < \mu(\delta), \qquad \|\boldsymbol{x}_{\mu(\Delta)}\| > \|\boldsymbol{x}_{\mu(\delta)}\|. \tag{2.7}$$

Further, let $\boldsymbol{r}_{\mu(\Delta)} = \widetilde{\boldsymbol{b}} - A\boldsymbol{x}_{\mu(\Delta)}$ and $\boldsymbol{r}_{\mu(\delta)} = \widetilde{\boldsymbol{b}} - A\boldsymbol{x}_{\mu(\delta)}$. Then $\|\boldsymbol{r}_{\mu(\Delta)}\| < \|\boldsymbol{r}_{\mu(\delta)}\|$. This situation is illustrated by Figure 2.1. In particular, $\boldsymbol{x}_{\mu(\Delta)}$ does not satisfy the discrepancy principle (1.8). □

Henceforth, we consider methods that do not make use of A^T for reasons outlined in Section 1, and we assume for notational simplicity that $A \in \mathbb{R}^{n\times n}$. Application of ℓ steps of the Arnoldi process with initial vector $\widetilde{\boldsymbol{b}}$ yields the Arnoldi decomposition

$$AV_\ell = V_{\ell+1}\bar{H}_\ell, \tag{2.8}$$

where $V_{\ell+1} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{\ell+1}] \in \mathbb{R}^{n\times(\ell+1)}$ has orthonormal columns, which span the Krylov subspace

$$\mathbb{K}_{\ell+1}(A, \widetilde{\boldsymbol{b}}) = \text{span}\{\widetilde{\boldsymbol{b}}, A\widetilde{\boldsymbol{b}}, \ldots, A^\ell\widetilde{\boldsymbol{b}}\},$$

and $\boldsymbol{v}_1 = \widetilde{\boldsymbol{b}}/\|\widetilde{\boldsymbol{b}}\|$. The matrix $V_\ell \in \mathbb{R}^{n\times\ell}$ is made up of the first ℓ columns of $V_{\ell+1}$. We assume that ℓ is chosen sufficiently small so that $\bar{H}_\ell \in \mathbb{R}^{(\ell+1)\times\ell}$ is an upper Hessenberg matrix with nonvanishing subdiagonal entries. In the rare event that for some $\ell \geq 1$ the last subdiagonal entry of $\bar{H}_\ell$ vanishes, the computations simplify. We will not dwell on this situation. Further details on the Arnoldi process can be found in, e.g., [27].

The range-restricted Arnoldi decomposition, as described in [18], is of the form

$$AV_\ell = W_{\ell+2}\bar{H}_\ell, \tag{2.9}$$

where the columns of $W_{\ell+2} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{\ell+2}] \in \mathbb{R}^{n\times(\ell+2)}$ form an orthonormal basis of $\mathbb{K}_{\ell+1}(A, \widetilde{\boldsymbol{b}})$ with $\boldsymbol{w}_1 = \widetilde{\boldsymbol{b}}/\|\widetilde{\boldsymbol{b}}\|$, the columns of $V_\ell \in \mathbb{R}^{n\times\ell}$ form an orthonormal basis of $\mathbb{K}_\ell(A, A\widetilde{\boldsymbol{b}})$, and $\bar{H}_\ell \in \mathbb{R}^{(\ell+2)\times\ell}$ vanishes below the sub-subdiagonal. Thus, $\mathcal{R}(V_\ell) \subset \mathcal{R}(A)$. Tikhonov regularization based on the range-restricted Arnoldi decomposition (2.9) tends to yield more accurate approximations of the desired solution $\widehat{\boldsymbol{x}}$ than Tikhonov regularization based on the standard Arnoldi decomposition (2.8) when the data $\widetilde{\boldsymbol{b}}$ is contaminated by noise. This is illustrated in Section 5. We remark that the decomposition (2.9) has better numerical properties than the range-restricted Arnoldi decomposition used in [15]. A comparison of these decompositions can be found in [18]. Typically, the parameter ℓ in the decompositions (2.8) and (2.9) is quite small and much smaller than n; see Section 5 for examples.

Let the matrix V_ℓ be defined by the decompositions (2.8) or (2.9). Substituting $\boldsymbol{x} = V_\ell\boldsymbol{y}$ into (1.5) and using (2.8) or (2.9) gives a minimization problem of the form

$$\min_{\boldsymbol{y}\in\mathbb{R}^\ell} \{\|\bar{H}_\ell\boldsymbol{y} - \boldsymbol{e}_1\|\widetilde{\boldsymbol{b}}\|\,\|^2 + \mu\,\|\boldsymbol{y}\|^2\}$$

with solution

$$\boldsymbol{y}_{\mu,\ell} = (\bar{H}_\ell^T\bar{H}_\ell + \mu I)^{-1}\bar{H}_\ell^T\boldsymbol{e}_1\|\widetilde{\boldsymbol{b}}\|,$$

where $\boldsymbol{e}_1 = [1, 0, \ldots, 0]^T \in \mathbb{R}^{k+1}$ denotes the first axis vector. Let

$$\boldsymbol{x}_{\mu,\ell} = V_\ell\boldsymbol{y}_{\mu,\ell} \tag{2.10}$$

and define the function

$$\psi_\ell(\mu) = \|\boldsymbol{x}_{\mu,\ell}\|^2, \qquad \mu > 0. \tag{2.11}$$

The following results are analogous to those of Propositions 2.2 and 2.3, and can be shown in similar ways.

Proposition 2.4. *Let $\bar{H}_\ell$ be defined by the Arnoldi decompositions (2.8) or (2.9), and assume that $\bar{H}_\ell^T \boldsymbol{e}_1 \neq \boldsymbol{0}$. Then the function (2.11) can be expressed as*

$$\psi_\ell(\mu) = \|\widetilde{\boldsymbol{b}}\|^2 \boldsymbol{e}_1^T \bar{H}_\ell (\bar{H}_\ell^T \bar{H}_\ell + \mu I)^{-2} \bar{H}_\ell^T \boldsymbol{e}_1,$$

which shows that $\psi_\ell(\mu)$ is strictly decreasing and convex for $\mu > 0$. Furthermore, the equation

$$\psi_\ell(\mu) = \tau$$

has a unique solution $0 < \mu < \infty$ for any $0 < \tau < \|\bar{H}_\ell^\dagger \boldsymbol{e}_1\|^2 \|\widetilde{\boldsymbol{b}}\|^2$.

Let $\boldsymbol{x}_{\mu,\ell}$ be given by (2.10) and introduce the function

$$\phi_\ell(\mu) = \|\widetilde{\boldsymbol{b}} - A\boldsymbol{x}_{\mu,\ell}\|^2, \qquad \mu > 0, \tag{2.12}$$

analogous to (2.4).

Proposition 2.5. *The function (2.12) can be expressed as*

$$\phi_\ell(\mu) = \|\widetilde{\boldsymbol{b}}\|^2 \boldsymbol{e}_1^T (\mu^{-1} \bar{H}_\ell \bar{H}_\ell^T + I)^{-2} \boldsymbol{e}_1,$$

where the matrix $\bar{H}_\ell$ is given by the Arnoldi decompositions (2.8) or (2.9). Assume that $\bar{H}_\ell^T \boldsymbol{e}_1 \neq \boldsymbol{0}$. Then $\phi_\ell(\mu)$ is strictly increasing for $\mu > 0$, and the equation

$$\phi_\ell(\mu) = \tau$$

has a unique solution $0 < \mu < \infty$ for any τ with

$$\|P_{\mathcal{N}(\bar{H}_\ell^T)} \boldsymbol{e}_1\|^2 \|\widetilde{\boldsymbol{b}}\|^2 < \tau < \|\widetilde{\boldsymbol{b}}\|^2,$$

where $P_{\mathcal{N}(\bar{H}_\ell^T)}$ denotes the orthogonal projector onto the null space of $\bar{H}_\ell^T$.

3. Combining solution norm and discrepancy constraints.

We consider the situation when both the norm of $\widehat{\boldsymbol{x}}$ and of the error $\widetilde{\boldsymbol{e}}$ are available and describe how this information can be applied when solving problems of small to medium size. As pointed out in Section 2, the discrepancy principle yields approximate solutions of norm smaller than (1.2). Moreover, the solution $\boldsymbol{x}_{\mu(\Delta)}$ of (1.4) with Δ defined by (1.2) satisfies $\|A\boldsymbol{x}_{\mu(\Delta)} - \widetilde{\boldsymbol{b}}\| < \eta\varepsilon$; see Example 2.1. However, the desired solution $\widehat{\boldsymbol{x}}$ does not satisfy this inequality. This indicates that $\boldsymbol{x}_{\mu(\Delta)}$ may be contaminated by propagated error.

The deficiencies of $\boldsymbol{x}_{\mu(\Delta)}$ and of the approximate solution determined with the aid of the discrepancy principle leads us to investigate whether requiring that the computed solution satisfies (1.8) and is of norm (1.2) can yield more accurate approximations of $\widehat{\boldsymbol{x}}$. Numerical examples reported in Section 5 show that this, indeed, can be the case.

Let $\boldsymbol{x}_{\rm d}$ satisfy (1.5) with the parameter $\mu > 0$ chosen so that $\boldsymbol{x}_{\rm d}$ satisfies (1.8), and solve the minimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \|\boldsymbol{x} - \boldsymbol{x}_{\rm d}\| \quad \text{with constraints} \quad \|\boldsymbol{x}\| = \Delta, \quad \|A\boldsymbol{x} - \widetilde{\boldsymbol{b}}\| = \eta\varepsilon. \tag{3.1}$$

The solution of (3.1) is of larger norm than $\boldsymbol{x}_{\mathrm{d}}$. Geometrically, we seek to determine a closest point to $\boldsymbol{x}_{\mathrm{d}}$ on the intersection of a sphere and an ellipsoid. Any solution is satisfactory.

Introduce the Lagrange function

$$\zeta_{\mu_1,\mu_2}(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_{\mathrm{d}}\|^2 + \mu_1\,(\|\boldsymbol{x}\|^2 - \Delta^2) + \mu_2\,(\|A\boldsymbol{x} - \widetilde{\boldsymbol{b}}\|^2 - \eta^2\varepsilon^2). \tag{3.2}$$

Differentiation with respect to $\boldsymbol{x}$, μ_1, and μ_2 yields the nonlinear system of equations for $\boldsymbol{x}$, μ_1, and μ_2:

$$\left\{\begin{array}{ccc} (\mu_2 A^T A + (\mu_1+1)I)\,\boldsymbol{x} & = & \boldsymbol{x}_{\mathrm{d}} + \mu_2 A^T\widetilde{\boldsymbol{b}}, \\ \|\boldsymbol{x}\|^2 & = & \Delta^2, \\ \|A\boldsymbol{x} - \widetilde{\boldsymbol{b}}\|^2 & = & \eta^2\varepsilon^2. \end{array}\right. \tag{3.3}$$

For small to medium-sized problems, we solve this system with the aid of the singular value decomposition

$$\begin{array}{lcll} A & = & \breve{U}\breve{\Sigma}\breve{V}^T, & \\ \breve{U} & = & [\breve{\boldsymbol{u}}_1, \breve{\boldsymbol{u}}_2, \ldots, \breve{\boldsymbol{u}}_m] \in \mathbb{R}^{m\times m}, & \breve{U}^T\breve{U} = I, \\ \breve{\Sigma} & = & \mathrm{diag}[\breve{\sigma}_1, \breve{\sigma}_2, \ldots, \breve{\sigma}_n], & \breve{\sigma}_1 \geq \breve{\sigma}_2 \geq \ldots \geq \breve{\sigma}_n \geq 0, \\ \breve{V} & = & [\breve{\boldsymbol{v}}_1, \breve{\boldsymbol{v}}_2, \ldots, \breve{\boldsymbol{v}}_n] \in \mathbb{R}^{n\times n}, & \breve{V}^T\breve{V} = I. \end{array} \tag{3.4}$$

Substituting this decomposition into (3.3) and letting $\boldsymbol{y} = \breve{V}^T\boldsymbol{x}$ yields

$$\left\{\begin{array}{ccc} (\mu_2\breve{\Sigma}^T\breve{\Sigma} + (\mu_1+1)I)\,\boldsymbol{y} & = & \breve{V}^T\boldsymbol{x}_{\mathrm{d}} + \mu_2\breve{\Sigma}^T\breve{U}^T\widetilde{\boldsymbol{b}}, \\ \|\boldsymbol{y}\|^2 & = & \Delta^2, \\ \|\breve{\Sigma}\boldsymbol{y} - U^T\widetilde{\boldsymbol{b}}\|^2 & = & \eta^2\varepsilon^2. \end{array}\right.$$

Introduce $\gamma_j = (\breve{U}^T\boldsymbol{b})_j$ and $\xi_j = (\breve{V}^T\boldsymbol{x}_{\mathrm{d}})_j$ for $j = 1, \ldots, n$. We are interested in computing a zero of the function

$$F(\mu_1,\mu_2) = \begin{bmatrix} \displaystyle\sum_{j=1}^{n}\left(\frac{\breve{\sigma}_j\gamma_j\mu_2 + \xi_j}{\mu_2\breve{\sigma}_j^2 + \mu_1 + 1}\right)^2 - \Delta^2 \\ \displaystyle\sum_{j=1}^{n}\left(\frac{\breve{\sigma}_j^2\gamma_j\mu_2 + \xi_j\breve{\sigma}_j}{\mu_2\breve{\sigma}_j^2 + \mu_1 + 1} - \gamma_j\right)^2 - \eta^2\varepsilon^2 \end{bmatrix}.$$

This may be done, e.g., by Newton's method.

Large-scale problems may be reduced by substituting one of the Arnoldi decompositions (2.8) or (2.9), or a partial Lanczos bidiagonalization of A, into (3.2). The reduced problem so obtained can be solved with the aid of the singular value decomposition as described above.

An alternative approach to combine the discrepancy principle and a solution norm constraint for large-scale problems is to use the Arnoldi method with solution norm constraint (as described in the previous section), and terminate the iterations with the Arnoldi method as soon as the discrepancy principle is satisfied. The performances of this approach, as well as of the other methods discussed in this and the previous sections, are illustrated in Section 5.

4. Sensitivity analysis. This section studies the sensitivity of the regularization parameter μ in (1.5) to changes in Δ, defined by (1.2), and to perturbations in the bound ε for the norm of the noise (1.7). This analysis is motivated by the fact that only approximations of the bound (1.7) and of (1.2) may be available.

Let $\delta = \eta\varepsilon$. It is convenient to let $\mu_{\rm d}$ denote the value of the regularization parameter for which (1.8) is satisfied and to define $\boldsymbol{x}_{\rm d} = \boldsymbol{x}_{\mu_{\rm d}}$. Similarly, let $\mu_{\rm n}$ be the value of the regularization parameter such that $\|\boldsymbol{x}_{\mu_{\rm n}}\| = \Delta$ and introduce $\boldsymbol{x}_{\rm n} = \boldsymbol{x}_{\mu_{\rm n}}$. Moreover, we denote the residual by

$$\boldsymbol{r} = \boldsymbol{b} - A\boldsymbol{x}\,;$$

in particular, $\boldsymbol{r}_{\rm d} = \boldsymbol{b} - A\boldsymbol{x}_{\rm d}$. Using this notation, the discussion following Proposition 2.3 can be summarized as

$$\mu_{\rm n} < \mu_{\rm d}, \quad \|\boldsymbol{x}_{\rm d}\| < \|\boldsymbol{x}_{\rm n}\| \quad \text{for } \delta > 0.$$

We are interested in the sensitivity of $\mu_{\rm n} = \mu_{\rm n}(\Delta)$ and $\mu_{\rm d} = \mu_{\rm d}(\delta)$ to perturbations in Δ and δ, respectively. The bounds below provide some insight.

PROPOSITION 4.1. *We have*

$$\frac{\mu_{\rm n}}{\Delta} \le |\mu_{\rm n}'(\Delta)| \le \frac{\|A\|^2 + \mu_{\rm n}}{\Delta} \tag{4.1}$$

and

$$\max\left\{\frac{\delta}{\|\boldsymbol{x}_{\rm d}\|^2},\ \frac{\delta\,\mu_{\rm d}^2}{\delta_-^2}\right\} \le\ \mu_{\rm d}'(\delta) \le \frac{\|A\|^2 + \mu_{\rm d}}{\mu_{\rm d}\,\|\boldsymbol{x}_{\rm d}\|^2}\,\delta, \tag{4.2}$$

where

$$\delta_-^2 = \sum_{j=1}^{r} \frac{\mu_{\rm d}^2}{(\breve{\sigma}_j^2 + \mu_{\rm d})^2}\,(\breve{\boldsymbol{u}}_j^T\widetilde{\boldsymbol{b}})^2$$

and r is the rank of A. Thus, $\delta_-^2 \le \delta^2$, with equality when A is square and nonsingular.

Proof. We first show the inequalities (4.1). For this purpose, we express the relation between $\mu_{\rm n}$ and Δ in terms of the singular value decomposition (3.4) and obtain

$$\|\boldsymbol{x}_{\rm n}\|^2 = \sum_{j=1}^{r} \frac{\breve{\sigma}_j^2}{(\breve{\sigma}_j^2 + \mu_{\rm n})^2}\,(\breve{\boldsymbol{u}}_j^T\widetilde{\boldsymbol{b}})^2 = \Delta^2. \tag{4.3}$$

Considering $\mu_{\rm n} = \mu_{\rm n}(\Delta)$ as a function of Δ and differentiating (4.3) with respect to Δ gives

$$\mu_{\rm n}'(\Delta) = -\Delta\left(\sum_{j=1}^{r} \frac{\breve{\sigma}_j^2}{(\breve{\sigma}_j^2 + \mu_{\rm n})^3}\,(\breve{\boldsymbol{u}}_j^T\widetilde{\boldsymbol{b}})^2\right)^{-1}. \tag{4.4}$$

Therefore, $\mu_{\rm n}'(\Delta) < 0$ and

$$|\mu_{\rm n}'(\Delta)| \le \Delta\,(\breve{\sigma}_1^2 + \mu_{\rm n})\left(\sum_{j=1}^{r} \frac{\breve{\sigma}_j^2}{(\breve{\sigma}_j^2 + \mu_{\rm n})^2}\,(\breve{\boldsymbol{u}}_j^T\widetilde{\boldsymbol{b}})^2\right)^{-1} = \frac{\|A\|^2 + \mu_{\rm n}}{\Delta}. \tag{4.5}$$

Moreover,

$$|\mu_{\rm n}'(\Delta)| \geq \Delta\,\mu_{\rm n} \left(\sum_{j=1}^{r} \frac{\breve{\sigma}_j^2}{(\breve{\sigma}_j^2+\mu_{\rm n})^2}\,(\breve{\boldsymbol{u}}_j^T\widetilde{\boldsymbol{b}})^2\right)^{-1} = \frac{\mu_{\rm n}}{\Delta}. \tag{4.6}$$

We turn to (4.2) and first show the lower bounds. The regularization parameter $\mu_{\rm d} = \mu_{\rm d}(\delta)$ is such that

$$\|\boldsymbol{r}_{\rm d}\|^2 = \sum_{j=1}^{r} \frac{\mu_{\rm d}^2}{(\breve{\sigma}_j^2+\mu_{\rm d})^2}\,(\breve{\boldsymbol{u}}_j^T\widetilde{\boldsymbol{b}})^2 + \sum_{j=r+1}^{m} (\breve{\boldsymbol{u}}_j^T\widetilde{\boldsymbol{b}})^2 = \delta^2, \tag{4.7}$$

where $\delta = \eta\varepsilon$. Differentiating (4.7) with respect to δ yields

$$\mu_{\rm d}'(\delta) = \frac{\delta}{\mu_{\rm d}} \left(\sum_{j=1}^{r} \frac{\breve{\sigma}_j^2}{(\breve{\sigma}_j^2+\mu_{\rm d})^3}\,(\breve{\boldsymbol{u}}_j^T\widetilde{\boldsymbol{b}})^2\right)^{-1}. \tag{4.8}$$

It follows from the inequality

$$\frac{\breve{\sigma}_j^2}{(\breve{\sigma}_j^2+\mu_{\rm d})^3} \leq \frac{1}{(\breve{\sigma}_j^2+\mu_{\rm d})^2}$$

that

$$\mu_{\rm d}'(\delta) \geq \delta\,\mu_{\rm d} \left(\sum_{j=1}^{r} \frac{\mu_{\rm d}^2}{(\breve{\sigma}_j^2+\mu_{\rm d})^2}\,(\breve{\boldsymbol{u}}_j^T\widetilde{\boldsymbol{b}})^2\right)^{-1} = \frac{\delta\,\mu_{\rm d}}{\delta_-^2}. \tag{4.9}$$

When, instead, substituting the inequality $\breve{\sigma}_j^2 + \mu_{\rm d} \geq \mu_{\rm d}$ into (4.8), we obtain

$$\mu_{\rm d}'(\delta) \geq \delta \left(\sum_{j=1}^{r} \frac{\breve{\sigma}_j^2}{(\breve{\sigma}_j^2+\mu_{\rm d})^2}\,(\breve{\boldsymbol{u}}_j^T\widetilde{\boldsymbol{b}})^2\right)^{-1} = \frac{\delta}{\|\boldsymbol{x}_{\rm d}\|^2}.$$

The upper bound of (4.2) follows by substituting

$$\frac{\breve{\sigma}_j^2}{(\breve{\sigma}_j^2+\mu_{\rm d})^3} \geq \frac{1}{\|A\|^2+\mu_{\rm d}}\,\frac{\breve{\sigma}_j^2}{(\breve{\sigma}_j^2+\mu_{\rm d})^2}$$

into (4.8). □

We remark that also other bounds than in Proposition 4.1 can be derived by analogous techniques. Elementary computations give the sensitivity of the solution norm and residual norm to perturbations in μ; cf. Propositions 2.2 and 2.3.

Corollary 4.2. *We have*

$$\frac{\Delta}{\|A\|^2+\mu_{\rm n}} \leq |\Delta'(\mu_{\rm n})| \leq \frac{\Delta}{\mu_{\rm n}}$$

and

$$\frac{\mu_{\rm d}\,\|\boldsymbol{x}_{\rm d}\|^2}{(\|A\|^2+\mu_{\rm d})\,\delta} \leq \delta'(\mu_{\rm d}) \leq \min\left\{\frac{\|\boldsymbol{x}_{\rm d}\|^2}{\delta},\ \frac{\delta_-^2}{\delta\,\mu_{\rm d}}\right\}.$$

Now let us study the effect of a perturbation of μ on the approximate solution $\boldsymbol{x}_\mu$ given by (1.6). Using the singular value decomposition (3.4) of A, we obtain

$$\boldsymbol{x}_\mu = \sum_{j=1}^{r} \frac{\breve{\sigma}_j}{\breve{\sigma}_j^2 + \mu} \, (\breve{\boldsymbol{u}}_j^T \widetilde{\boldsymbol{b}}) \breve{\boldsymbol{v}}_j,$$

which shows that

$$\boldsymbol{x}_{\widetilde{\mu}} - \boldsymbol{x}_\mu = (\mu - \widetilde{\mu}) \sum_{j=1}^{r} \frac{\breve{\sigma}_j}{(\breve{\sigma}_j^2 + \mu)^2} \, (\breve{\boldsymbol{u}}_j^T \widetilde{\boldsymbol{b}}) \breve{\boldsymbol{v}}_j + \mathcal{O}((\mu - \widetilde{\mu})^2),$$

where we have assumed that $|\mu - \widetilde{\mu}| \ll \mu$. Therefore,

$$\| \boldsymbol{x}_{\widetilde{\mu}} - \boldsymbol{x}_\mu \| \leq \frac{|\mu - \widetilde{\mu}|}{\mu} \, \| \boldsymbol{x}_\mu \| + \mathcal{O}((\mu - \widetilde{\mu})^2).$$

Now applying the triangle inequality,

$$| \, \| \boldsymbol{x}_{\widetilde{\mu}} \| - \| \boldsymbol{x}_\mu \| \, | \leq \| \boldsymbol{x}_{\widetilde{\mu}} - \boldsymbol{x}_\mu \|,$$

gives the following results.

Proposition 4.3. *Let $\mu > 0$. Then*

$$\frac{d}{d\mu} \, \| \boldsymbol{x}_\mu \| \leq \frac{\| \boldsymbol{x}_\mu \|}{\mu}.$$

Corollary 4.4. *Let $\mu = \mu(\beta) > 0$ be a continuously differentiable function of the parameter β, and denote $\mu_0 = \mu(\beta_0)$. Then*

$$\lim_{\beta \to \beta_0} \frac{\| \boldsymbol{x}_{\mu(\beta)} - \boldsymbol{x}_{\mu_0} \|}{|\beta - \beta_0| \, \| \boldsymbol{x}_{\mu_0} \|} \leq \frac{|\mu'(\beta_0)|}{\mu_0}.$$

Corollary 4.4 in combination with Proposition 4.1 may be used to provide sensitivity bounds for $\boldsymbol{x}_\mu$ for the Tikhonov approaches based on the discrepancy principle and solution norm constraint. From (2.7) we know that $\mu_{\rm d} > \mu_{\rm n}$ for $\delta > 0$; in experiments in Section 5, the ratio $\mu_{\rm d}/\mu_{\rm n}$ was typically between 3 and 100. On the other hand, assuming modest (approximately $\mathcal{O}(1)$) values for $\|A\|$, $\|\Delta\|$, $\|\delta\|$, and $\|\boldsymbol{x}_{\rm d}\|$, both the upper and lower bounds for $\mu'(\Delta)$ in Proposition 4.1 generally will be smaller than those for $\mu'(\delta)$. We will show a related experiment in the following section.

5. Numerical experiments. We will provide several examples of the behavior of the various methods described in this paper and compare the results with known approaches. All our test problems are from Hansen's Regularization Tools [11]. The matrices in all examples are square, i.e., $m = n$. Unless stated otherwise, we use the following parameters in the examples: $\varepsilon = 0.01 \, \|\boldsymbol{b}\|$ in (1.7) (corresponding to 1% noise), and $\eta = 1.01$ in (1.8). As a measure of the quality of the approximations we tabulate the relative error $\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\| / \|\widehat{\boldsymbol{x}}\|$. Subsections 5.1-5.3 consider problems of small size, which we solve with the aid of the SVD of A.

5.1. A comparison of three methods for small-scale examples. We first consider small-scale examples ($n = 100$), and compare in Table 5.1 the qualities (relative errors) of the approximate solutions given by three approaches:

- Tikhonov regularization with the discrepancy principle (columns 2 and 5);
- Tikhonov regularization with solution norm constraint (columns 4 and 7);
- and the combination of discrepancy principle and solution norm constraint; see (3.3) (columns 3 and 6).

Two noise levels are considered: 1% and 10%. For the lower noise level, the entries of the columns "Tikhonov" and "$\|\boldsymbol{x}\|$" behave as one may expect. The solutions determined with the solution norm constraint are of larger norm than the solutions obtained with the discrepancy principle. The convergence property (1.9) of solutions determined with the discrepancy principle leads us to expect that the discrepancy principle will often yield more accurate approximations of $\widehat{\boldsymbol{x}}$ than the solution norm constraint. A comparison of columns 2 and 4 of Table 5.1 shows that this is indeed the case, although not for all test problems; see also Section 5.2. The third column illustrates that for most problems further accuracy can be achieved by applying both the discrepancy principle and the solution norm constraint as in (3.3).

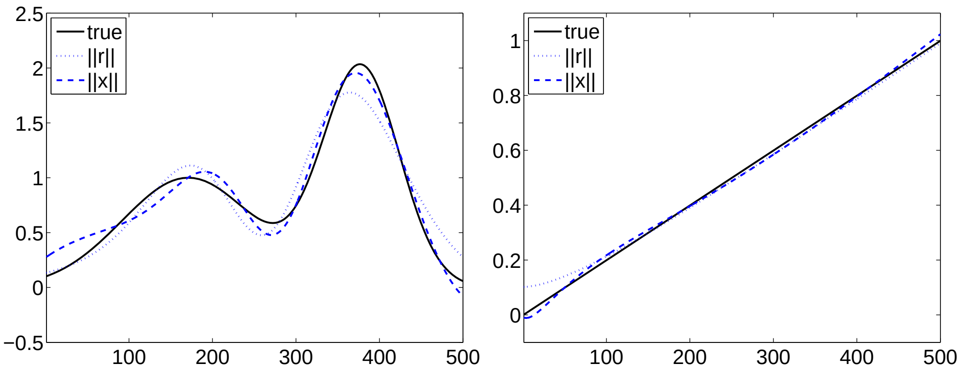
The situation changes when the noise in $\widetilde{\boldsymbol{b}}$ is increased to 10%. Now the discrepancy principle gives higher accuracy than the solution norm constraint in only half the experiments. Thus, for large noise levels the use of the solution norm constraint can be effective. For a few problems the best approximation of $\widehat{\boldsymbol{x}}$ is obtained by applying both the discrepancy principle and the solution norm constraint as in (3.3).

Table 5.1

Comparison of Tikhonov regularization based on the discrepancy principle ("Tikhonov"), Tikhonov regularization with solution norm constraint ("$\|\boldsymbol{x}\|$"), and the combination technique of (3.3)*, for* $n = 100$ *examples with 1% (columns 2–4) and 10% (columns 5–7) noise, respectively.*

		1% noise			10% noise	
Problem	Tikhonov	(3.3)	$\|\boldsymbol{x}\|$	Tikhonov	(3.3)	$\|\boldsymbol{x}\|$
baart	$1.68 \cdot 10^{-1}$	$1.63 \cdot 10^{-1}$	$6.19 \cdot 10^{-2}$	$3.01 \cdot 10^{-1}$	$2.48 \cdot 10^{-1}$	$1.50 \cdot 10^{-1}$
deriv2-1	$2.55 \cdot 10^{-1}$	$2.60 \cdot 10^{-1}$	$3.18 \cdot 10^{-1}$	$3.68 \cdot 10^{-1}$	$3.38 \cdot 10^{-1}$	$4.24 \cdot 10^{-1}$
deriv2-2	$2.41 \cdot 10^{-1}$	$2.43 \cdot 10^{-1}$	$2.97 \cdot 10^{-1}$	$3.73 \cdot 10^{-1}$	$3.30 \cdot 10^{-1}$	$4.30 \cdot 10^{-1}$
deriv2-3	$2.96 \cdot 10^{-2}$	$2.96 \cdot 10^{-2}$	$4.17 \cdot 10^{-2}$	$5.17 \cdot 10^{-2}$	$6.09 \cdot 10^{-2}$	$9.51 \cdot 10^{-2}$
foxgood	$3.27 \cdot 10^{-2}$	$3.21 \cdot 10^{-2}$	$3.41 \cdot 10^{-2}$	$7.65 \cdot 10^{-2}$	$6.17 \cdot 10^{-2}$	$4.01 \cdot 10^{-2}$
gravity	$2.35 \cdot 10^{-2}$	$2.34 \cdot 10^{-2}$	$2.04 \cdot 10^{-2}$	$6.83 \cdot 10^{-2}$	$7.01 \cdot 10^{-2}$	$6.79 \cdot 10^{-2}$
heat	$1.46 \cdot 10^{-1}$	$1.40 \cdot 10^{-1}$	$2.02 \cdot 10^{-1}$	$3.73 \cdot 10^{-1}$	$3.21 \cdot 10^{-1}$	$4.65 \cdot 10^{-1}$
ilaplace	$1.47 \cdot 10^{-1}$	$1.34 \cdot 10^{-1}$	$1.76 \cdot 10^{-1}$	$1.99 \cdot 10^{-1}$	$1.99 \cdot 10^{-1}$	$1.93 \cdot 10^{-1}$
phillips	$2.90 \cdot 10^{-2}$	$2.90 \cdot 10^{-2}$	$5.49 \cdot 10^{-2}$	$6.91 \cdot 10^{-2}$	$1.01 \cdot 10^{-1}$	$1.50 \cdot 10^{-1}$
shaw	$1.32 \cdot 10^{-1}$	$8.80 \cdot 10^{-2}$	$1.08 \cdot 10^{-1}$	$1.72 \cdot 10^{-1}$	$1.77 \cdot 10^{-1}$	$1.67 \cdot 10^{-1}$

In Figure 5.1 we consider two specific examples of size $n = 500$, 1% noise, and $\eta = 1.1$ (in contrast to $\eta = 1.01$ in Table 5.1). Figure 5.1(a) shows shaw: true solution (solid line), Tikhonov regularization matching the discrepancy principle (relative error 0.15; dotted graph), and Tikhonov regularization with solution norm constraint $\|\boldsymbol{x}\| = \|\widehat{\boldsymbol{x}}\|$ (relative error 0.096; dashed graph). Thus, the solution norm constraint gives higher accuracy than the discrepancy principle. It was the other way in Table 5.1. The significance of the noise level and the parameter η are further illustrated in following subsections. We note that truncated singular value decomposition (TSVD), with the truncation index k chosen to be as large as possible so that the computed approximate solution, $\boldsymbol{x}_k$, satisfies the discrepancy principle $\|A\boldsymbol{x}_k - \widetilde{\boldsymbol{b}}\| \leq \eta\varepsilon$, yields the relative error $\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}\|/\|\hat{\boldsymbol{x}}\| = 0.17$. This error is larger than for Tikhonov regularization.



(a) (b)

FIG. 5.1. *Two $n = 500$ examples with 1% noise. (a)* shaw*: true solution (solid), Tikhonov based on the discrepancy principle $\|\boldsymbol{r}\| = \eta\varepsilon\|\boldsymbol{b}\|$ (dotted graph), Tikhonov based on a norm solution constraint $\|\boldsymbol{x}\| = \Delta$ (dashed graph). (b) Same for* foxgood*.*

Figure 5.1(b) displays foxgood. Here, Tikhonov regularization with the discrepancy principle yields the relative error 0.044, while Tikhonov regularization with the (exact) norm constraint gives the relative error 0.022. The TSVD method yields an approximate solution with relative error 0.031. Similarly as in Table 5.1, the solution norm constraint gives higher accuracy than the discrepancy principle.

5.2. The influence of the noise level. Next, we compare for various noise levels the quality of approximate solutions determined by Tikhonov regularization based on the discrepancy principle and Tikhonov regularization with solution norm constraint. In Figure 5.2 we plot the relative error of the approximations obtained with the discrepancy principle ($\|\widetilde{\boldsymbol{b}} - A\boldsymbol{x}\| = 1.1 \cdot \varepsilon$, where ε varies from $10^{-4}\|\boldsymbol{b}\|$ (very little noise) to $10^{-1}\|\boldsymbol{b}\|$ (much noise); marked in the figure by "$\|\boldsymbol{r}\|$") and the relative error of the approximations obtained with a solution norm constraint ($\|\boldsymbol{x}\| = \|\widehat{\boldsymbol{x}}\|$, marked by "$\|\boldsymbol{x}\|$"). We consider four different test problems of dimension $n = 500$.

Figure 5.2 shows the discrepancy principle to yield computed solutions of foxgood and gravity that approximate $\widehat{\boldsymbol{x}}$ more accurately than approximate solutions determined with the solution norm constraint when there is little noise. However, this is not the case for deriv2-1 and phillips. We conclude that imposing a solution norm constraint may be a valuable alternative to Tikhonov regularization based on the discrepancy principle.

5.3. Sensitivities as function of Δ and δ. We return to the situation of 1% noise, and study what happens for both Tikhonov regularization methods if the estimates concerning the residual norm or solution norm are inaccurate. For the discrepancy principle, we impose the requirement $\|\boldsymbol{r}\| = \eta\varepsilon$ for $\varepsilon = 0.01\|\boldsymbol{b}\|$ and varying η. The cases $\eta < 1$ and $\eta > 1$ may be viewed as underestimation and overestimation of the noise, respectively. For the solution norm approach, we use the estimate $\|\boldsymbol{x}\| = \eta\,\|\widehat{\boldsymbol{x}}\|$. Here, $\eta < 1$ and $\eta > 1$ may be seen as underestimation and overestimation of the norm of the true solution, respectively.

In Figure 5.3 we let η vary from 0.1 to 10 for two of the examples of Figure 5.2. As we clearly see, both methods perform the best for η close to unity. For the approach based on the solution norm constraint, it seems important that $\|\boldsymbol{x}\|$ not be underestimated. However, if both $\|\boldsymbol{x}\|$ and $\|\boldsymbol{r}\|$ (the discrepancy principle) are overestimated,

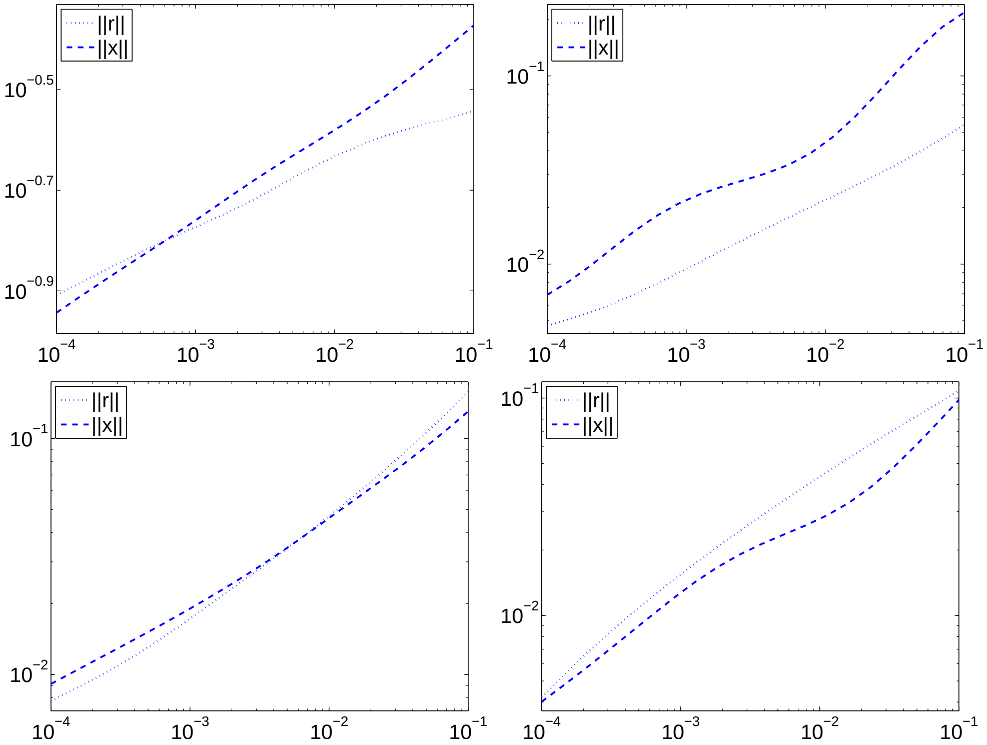
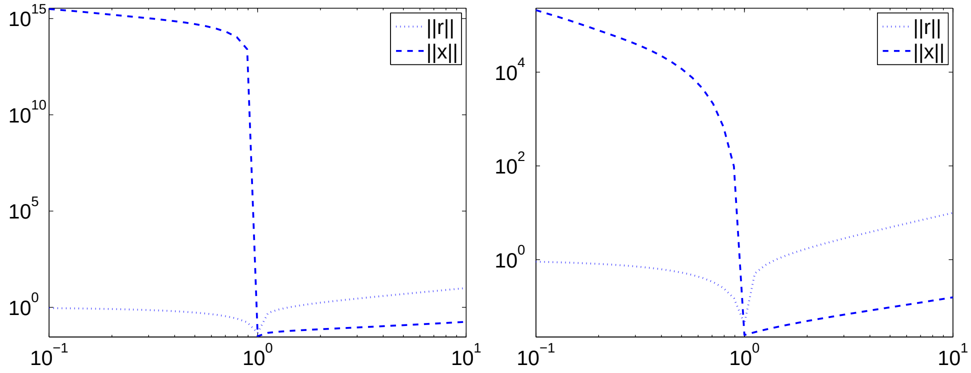


Fig. 5.2. *The qualities (relative errors) of Tikhonov regularization based on the discrepancy principle (dotted graph) versus Tikhonov based on a solution norm constraint (dashed graph) for* 500×500 *examples* deriv2-1 *(top-left),* foxgood *(top-right),* gravity *(bottom-left), and* phillips *(bottom-right) for various noise levels.*



(a) (b)

Fig. 5.3. *The qualities (relative errors) of Tikhonov based on the discrepancy principle (dot) versus Tikhonov based on a solution norm constraint (dash) for* gravity *(a) and* phillips *(b) for 1% noise and various qualities of the residual norm or solution norm estimate (η between 0.1 and 10, corresponding to underestimations and overestimations, respectively).*

then the method based on the solution norm constraint is clearly superior. This implies that the quality of the computed approximate solutions, when using the solution norm constraint, may be fairly insensitive to incorrect estimates of the solution norm, as long as this estimate is larger than the norm of the true solution. We recall from Section 2 that an approximate solution determined with Δ larger than (1.2) also can be computed by imposing the discrepancy principle (1.8) for some $\eta > 0$.

5.4. Noise-free problems: solution norm matching Arnoldi–Tikhonov versus LSTRS and generalized Arnoldi. In this subsection we use the norm-matching Arnoldi–Tikhonov method based on the standard Arnoldi decomposition (2.8) to solve noise-free problems. We make a comparison with results reported by Lampe et al. [13] for the LSTRS method. Table 5.2 shows the relative error in the computed approximate solutions and the number of matrix-vector multiplications (MV) for various test problems considered in [13]. The parameter ℓ in the decomposition (2.8) is one smaller than the number of matrix-vector multiplications. The norm-matching Arnoldi–Tikhonov method matches the norm $\|\boldsymbol{x}_\ell\| = \|\widehat{\boldsymbol{x}}\|$ for increasing values of ℓ until the relative change in $\boldsymbol{x}_\ell$ or in μ_ℓ is less than 10^{-4}.

For the new method we test two approaches: $\varepsilon = 0$ in (1.7), which corresponds to no noise. As we see, the norm-matching Arnoldi–Tikhonov is superior to LSTRS with the exception of the heat and deriv2-2 examples. The method does not converge for the latter case since the norms of the rendered solutions in each iteration are too small. Therefore, as an alternative, we also give the performance of the method when we pretend that there is relative noise of level 10^{-6} in the right-hand side, i.e., we set $\varepsilon = 10^{-6}$ in (1.7) but let $\boldsymbol{b}$ be error-free. The method then terminates when the above mentioned criteria or the discrepancy principle are satisfied. This reduces the number of iterations. It may or may not improve the accuracy in the computed solution, but the results are again better than for LSTRS apart from the heat examples.

TABLE 5.2

Norm-matching Arnoldi–Tikhonov compared to LSTRS; with noise-free data $\widetilde{\boldsymbol{b}}$, for $n = 1000$. The last two columns are taken from [13].

	$\|\boldsymbol{x}\|,\ \varepsilon = 0$		$\|\boldsymbol{x}\|,\ \varepsilon = 10^{-6}$		LSTRS	
Problem	quality	MV	quality	MV		
baart	$2.8 \cdot 10^{-5}$	8	$1.5 \cdot 10^{-5}$	7	$8.6 \cdot 10^{-2}$	18
deriv2-1	$3.9 \cdot 10^{-8}$	161	$5.7 \cdot 10^{-2}$	37	$5.8 \cdot 10^{-1}$	217
deriv2-2	–	–	$5.5 \cdot 10^{-2}$	36	$3.4 \cdot 10^{-1}$	148
foxgood	$2.6 \cdot 10^{-4}$	7	$8.6 \cdot 10^{-4}$	6	$3.7 \cdot 10^{-2}$	18
heat ($\kappa = 5$)	$1.7 \cdot 10^{-2}$	61	$1.7 \cdot 10^{-2}$	61	$5.0 \cdot 10^{-3}$	68
heat ($\kappa = 1$)	$3.9 \cdot 10^{-1}$	88	$1.0 \cdot 10^{0}$	40	$8.1 \cdot 10^{-3}$	112
ilaplace-1	$1.3 \cdot 10^{-1}$	76	$2.2 \cdot 10^{-1}$	21	$3.3 \cdot 10^{-1}$	137
ilaplace-3	$1.5 \cdot 10^{-3}$	35	$4.7 \cdot 10^{-3}$	30	$6.7 \cdot 10^{-2}$	52
phillips	$2.7 \cdot 10^{-3}$	10	$1.2 \cdot 10^{-3}$	17	$9.9 \cdot 10^{-3}$	92
shaw	$5.9 \cdot 10^{-5}$	21	$3.2 \cdot 10^{-2}$	10	$5.9 \cdot 10^{-2}$	36

Results reported in [16, Table 2] for the generalized Arnoldi method make it possible to compare this method with Arnoldi–Tikhonov for heat(1000), phillips(1000), and shaw(1000). The generalized Arnoldi method performs better for heat(1000), but not for the other problems.

5.5. Arnoldi–Tikhonov: solution norm matching vs. the discrepancy principle. We turn to experiments with Arnoldi–Tikhonov methods when the data are noisy. Two different situations for solution norm matching Arnoldi–Tikhonov are considered. First, we assume that there is a bound (1.7) so that we also can use the discrepancy principle (1.8). Tables 5.3 and 5.4 show the results for various test problems of dimension $n = 1000$. We test two flavors: the standard (columns 2–3) and the range-restricted Arnoldi (columns 4–5) methods based on the decompositions (2.8) and (2.9), respectively. These norm-matching Arnoldi–Tikhonov methods match the norm $\|\boldsymbol{x}_\ell\| = \|\widehat{\boldsymbol{x}}\|$ for increasing values of the parameter ℓ in (2.8) and (2.9).[1] The computations are terminated if, additionally, the discrepancy principle (1.8) also is satisfied.

Columns 6–9 show the performance of the standard and range-restricted Arnoldi–Tikhonov methods. The computations are terminated as soon as the parameter ℓ in (2.8) and (2.9) is large enough so that the discrepancy principle can be satisfied. Tables 5.3 and 5.4 differ in the noise level (1% and 10%, respectively).

Table 5.3

Columns 2–5: norm-matching Arnoldi–Tikhonov (stopping if the discrepancy principle is satisfied) for $n = 1000$ examples with 1% noise in the right-hand side. Columns 6–9: Arnoldi–Tikhonov based on the discrepancy principle.

	$\|\boldsymbol{x}\|,\ \mathcal{K}(A,\boldsymbol{b})$		$\|\boldsymbol{x}\|,\ \mathcal{K}(A,A\boldsymbol{b})$		$\|\boldsymbol{r}\|,\ \mathcal{K}(A,\boldsymbol{b})$		$\|\boldsymbol{r}\|,\ \mathcal{K}(A,A\boldsymbol{b})$	
Problem	quality	MV	quality	MV	quality	MV	quality	MV
baart	$3.4 \cdot 10^{-2}$	4	$3.3 \cdot 10^{-2}$	3	$2.9 \cdot 10^{-1}$	3	$5.3 \cdot 10^{-2}$	3
deriv2-1	$3.7 \cdot 10^{-1}$	6	$2.7 \cdot 10^{-1}$	10	$4.3 \cdot 10^{-1}$	5	$2.4 \cdot 10^{-1}$	6
deriv2-2	$3.5 \cdot 10^{-1}$	6	$2.3 \cdot 10^{-1}$	8	$4.2 \cdot 10^{-1}$	5	$2.2 \cdot 10^{-1}$	6
deriv2-3	$4.7 \cdot 10^{-2}$	4	$2.0 \cdot 10^{-2}$	4	$9.8 \cdot 10^{-2}$	2	$2.6 \cdot 10^{-2}$	3
foxgood	$7.6 \cdot 10^{-2}$	2	$2.9 \cdot 10^{-2}$	4	$7.6 \cdot 10^{-2}$	2	$3.3 \cdot 10^{-2}$	2
gravity	$4.8 \cdot 10^{-2}$	7	$3.4 \cdot 10^{-2}$	8	$1.4 \cdot 10^{-1}$	5	$2.9 \cdot 10^{-2}$	6
heat	$1.6 \cdot 10^{-1}$	120	$1.6 \cdot 10^{-1}$	256	$7.2 \cdot 10^{8}$	63	$2.8 \cdot 10^{10}$	91
ilaplace	$2.5 \cdot 10^{-1}$	11	$2.5 \cdot 10^{-1}$	10	$1.7 \cdot 10^{0}$	7	$1.6 \cdot 10^{0}$	8
phillips	$3.2 \cdot 10^{-2}$	5	$3.4 \cdot 10^{-2}$	8	$9.6 \cdot 10^{-2}$	4	$2.5 \cdot 10^{-2}$	4
shaw	$1.5 \cdot 10^{-1}$	6	$5.7 \cdot 10^{-2}$	6	$1.1 \cdot 10^{-1}$	6	$1.1 \cdot 10^{-1}$	5

Table 5.4

Columns 2–5: norm-matching Arnoldi–Tikhonov (stopping if the discrepancy principle is satisfied) for $n = 1000$ examples with 10% noise in the right-hand side (standard and range-restricted Arnoldi). Columns 6–9: Arnoldi–Tikhonov based on the discrepancy principle.

	$\|\boldsymbol{x}\|,\ \mathcal{K}(A,\boldsymbol{b})$		$\|\boldsymbol{x}\|,\ \mathcal{K}(A,A\boldsymbol{b})$		$\|\boldsymbol{r}\|,\ \mathcal{K}(A,\boldsymbol{b})$		$\|\boldsymbol{r}\|,\ \mathcal{K}(A,A\boldsymbol{b})$	
Problem	quality	MV	quality	MV	quality	MV	quality	MV
baart	$5.7 \cdot 10^{-1}$	3	$3.2 \cdot 10^{-1}$	4	$5.0 \cdot 10^{-1}$	3	$5.1 \cdot 10^{-1}$	2
deriv2-1	$5.2 \cdot 10^{-1}$	4	$4.1 \cdot 10^{-1}$	6	$7.1 \cdot 10^{-1}$	3	$3.8 \cdot 10^{-1}$	3
deriv2-2	$4.7 \cdot 10^{-1}$	4	$3.4 \cdot 10^{-1}$	5	$7.6 \cdot 10^{-1}$	3	$3.5 \cdot 10^{-1}$	3
deriv2-3	$1.7 \cdot 10^{-1}$	2	$4.9 \cdot 10^{-2}$	3	$1.2 \cdot 10^{-1}$	2	$6.7 \cdot 10^{-2}$	2
foxgood	$1.1 \cdot 10^{-1}$	3	$8.6 \cdot 10^{-2}$	3	$4.3 \cdot 10^{-1}$	2	$1.2 \cdot 10^{-1}$	2
gravity	$1.9 \cdot 10^{-1}$	4	$1.1 \cdot 10^{-1}$	6	$3.8 \cdot 10^{-1}$	3	$7.7 \cdot 10^{-2}$	4
heat	$5.1 \cdot 10^{-1}$	58	$5.1 \cdot 10^{-1}$	91	$6.7 \cdot 10^{7}$	38	$1.4 \cdot 10^{7}$	54
ilaplace	$2.3 \cdot 10^{-1}$	9	$2.6 \cdot 10^{-1}$	9	$2.3 \cdot 10^{0}$	6	$1.8 \cdot 10^{0}$	6
phillips	$1.3 \cdot 10^{-1}$	4	$3.8 \cdot 10^{-2}$	4	$3.5 \cdot 10^{-1}$	3	$8.4 \cdot 10^{-2}$	3
shaw	$2.2 \cdot 10^{-1}$	5	$1.1 \cdot 10^{-1}$	5	$3.2 \cdot 10^{-1}$	4	$1.7 \cdot 10^{-1}$	4

The conclusion here is that the solution norm matching Arnoldi–Tikhonov method performs better than the Arnoldi–Tikhonov method based on the discrepancy principle for many of the test problems. As one specific example, we show the approximate

[1] In this section, we refer to the computed solution (2.10) as $\boldsymbol{x}_\ell$.

solution given by the Arnoldi–Tikhonov method with solution norm constraint for baart with 1% noise in Figure 5.4. The method stops after 4 iterations when $\|\boldsymbol{r}\| \le \eta\varepsilon$; the relative error in $\boldsymbol{x}$ is 0.034 (cf. the top-left entry of Table 5.3).

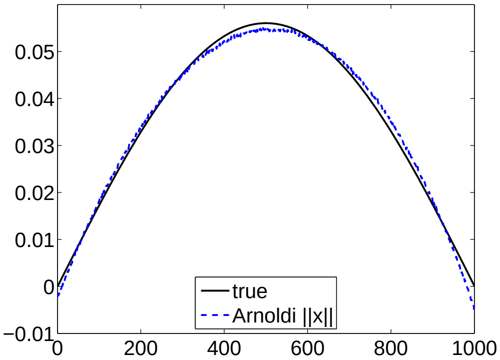


Fig. 5.4. *Example* baart*, $n = 1000$, 1% noise. True solution (solid graph) and Arnoldi–Tikhonov solution based on a solution norm constraint (dashed graph).*

Now assume instead that we have an estimate for the solution norm but that an error bound (1.7) is not available. In this situation, methods based on the discrepancy principle cannot be applied. In Table 5.5, we give the results for the Arnoldi–Tikhonov approach that satisfies the solution norm constraint for increasing values of the parameter ℓ in (2.8) and (2.9). The computations are terminated as soon as two consecutive approximations $\boldsymbol{x}_\ell$ or μ_ℓ differ by less than 10^{-4} relatively (the same stopping criterion as for the noise-free examples in Table 5.2). We see that for several test problems satisfactory approximations are computed without the knowledge of a bound for the norm of the noise (1.7) (and, consequently, without the use of the discrepancy principle).

Table 5.5

Norm-matching Arnoldi–Tikhonov for $n = 1000$ examples without use of the discrepancy principle (stopping if two consecutive approximations $\boldsymbol{x}_\ell$ or μ_ℓ differ by less than 10^{-4} relatively) for $n = 1000$ examples with 1% noise in the right-hand side.

	$\|\boldsymbol{x}\|$, $\mathcal{K}(A, \boldsymbol{b})$		$\|\boldsymbol{x}\|$, $\mathcal{K}(A, A\boldsymbol{b})$	
Problem	quality	MV	quality	MV
baart	$2.1 \cdot 10^{-1}$	13	$2.1 \cdot 10^{-1}$	12
deriv2-1	$2.8 \cdot 10^{-1}$	18	$2.8 \cdot 10^{-1}$	17
deriv2-2	$2.5 \cdot 10^{-1}$	16	$2.5 \cdot 10^{-1}$	15
deriv2-3	$3.0 \cdot 10^{-2}$	9	$3.0 \cdot 10^{-2}$	8
foxgood	$2.9 \cdot 10^{-2}$	7	$2.9 \cdot 10^{-2}$	6
gravity	$3.6 \cdot 10^{-2}$	11	$3.6 \cdot 10^{-2}$	10
heat	$7.3 \cdot 10^{-1}$	50	$7.6 \cdot 10^{-1}$	79
ilaplace	$2.5 \cdot 10^{-1}$	45	$2.5 \cdot 10^{-1}$	42
phillips	$4.2 \cdot 10^{-2}$	14	$4.2 \cdot 10^{-2}$	13
shaw	$5.4 \cdot 10^{-2}$	10	$5.4 \cdot 10^{-2}$	9

6. Conclusions.

We have presented several approaches that exploit a solution norm constraint. For small-scale problems we described a solution norm matching Tikhonov-type method, as well as a technique that yields an approximate solution that satisfies both a solution norm constraint and the discrepancy principle. We also discussed an Arnoldi–Tikhonov-type technique for large-scale problems. Our numerical examples lead us to the following observations:

- For some small-scale problems, the solution norm constraint may yield more accurate approximate solutions than the discrepancy principle.
- If it is important that the computed solution be of a particular norm and the discrepancy principle can be applied, then the methods of Section 3 may be attractive.
- Arnoldi–Tikhonov with a solution norm constraint may be used for noise-free and noise-contaminated problems, with and without the use of the discrepancy principle.
- Arnoldi–Tikhonov with a solution norm constraint performs better than the other methods in our comparison for many noise-free problems.
- Arnoldi–Tikhonov using both a solution norm constraint and the discrepancy principle yields more accurate approximate solutions than when only the discrepancy principle is applied.

In summary, methods that use a solution norm constraint may be helpful for computing accurate approximate solutions. The numerical examples show the use of both a solution norm constraint and the discrepancy principle to yield particularly accurate approximations of the desired solution.

Acknowledgments: We thank a referee for very helpful suggestions.

REFERENCES

[1] M. L. Baart, *The use of auto-correlation for pseudo-rank determination in noisy ill-conditioned least-squares problems*, IMA J. Numer. Anal., 2 (1982), pp. 241–247.

[2] D. Calvetti, B. Lewis, and L. Reichel, *On the choice of subspace for iterative methods for linear discrete ill-posed problems*, Int. J. Appl. Math. Comput. Sci., 11 (2001), pp. 1069–1092.

[3] D. Calvetti, B. Lewis, L. Reichel, and F. Sgallari, *Tikhonov regularization with nonnegativity constraint*, Electron. Trans. Numer. Anal., 18 (2004), pp. 153–173.

[4] D. Calvetti and L. Reichel, *Tikhonov regularization with a solution constraint*, SIAM J. Sci. Comput., 26 (2004), pp. 224–239.

[5] T. F. Chan and K. R. Jackson, *Nonlinearly preconditioned Krylov subspace methods for discrete Newton algorithms*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 533–542.

[6] L. Eldén, *A weighted pseudoinverse, generalized singular values, and constrained least squares problems*, BIT, 22 (1982), pp. 487–501.

[7] W. Gander, *Least squares with a quadratic constraint*, Numer. Math., 36 (1991), pp. 291–307.

[8] G. H. Golub and U. von Matt, *Quadratically constrained least squares and quadratic problems*, Numer. Math., 59 (1991), pp. 561–580.

[9] C. W. Groetsch, The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind, Pitman, Boston, 1984.

[10] P. C. Hansen, Rank-Deficient and Discrete Ill-Posed Problems, SIAM, Philadelphia, 1998.

[11] P. C. Hansen, *Regularization tools version 4.0 for Matlab 7.3*, Numer. Algorithms, 46 (2007), pp. 189–194.

[12] A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, Springer, New York, 1996.

[13] J. Lampe, M. Rojas, D. Sorensen, and H. Voss, *Accelerating the LSTRS algorithm*, SIAM J. Sci. Comput., 33 (2011), pp. 175–194.

[14] J. Lampe and H. Voss, *A fast algorithm for solving regularized total least squares problems*, Electron. Trans. Numer. Anal., 31 (2008), pp. 12–24.

[15] B. Lewis and L. Reichel, *Arnoldi–Tikhonov regularization methods*, J. Comput. Appl. Math., 226 (2009), pp. 92–102.

[16] R.-C. Li and Q. Ye, *A Krylov subspace method for quadratic matrix polynomials with application to constrained least squares problems*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 405–428.

[17] S. Morigi, L. Reichel, and F. Sgallari, *Orthogonal projection regularization operators*, Numer. Algorithms, 44 (2007), pp. 99–114.

[18] A. Neuman, L. Reichel, and H. Sadok, *Implementations of range restricted iterative methods for linear discrete ill-posed problems*, Linear Algebra Appl., in press.

[19] D. L. Phillips, *A technique for the numerical solution of certain integral equations of the first kind*, J. ACM, 9 (1962), pp. 84–97.

[20] L. Reichel and Q. Ye, *Simple square smoothing regularization operators*, Electron. Trans. Numer. Anal., 33 (2009), pp. 63–83.

[21] C. H. Reinsch, *Smoothing by spline functions*, Numer. Math., 10 (1967), pp. 177–183.

[22] C. H. Reinsch, *Smoothing by spline functions II*, Numer. Math., 16 (1971), pp. 451–454.

[23] M. Rojas, S. A. Santos, and D. C. Sorensen, *A new matrix-free algorithm for the large-scale trust-region subproblem*, SIAM J. Optim., 11 (2000), pp. 611–646.

[24] M. Rojas and D. C. Sorensen, *A trust-region approach to regularization of large-scale discrete forms of ill-posed problems*, SIAM J. Sci. Comput., 23 (2002), pp. 1842–1860.

[25] M. Rojas and T. Steihaug, *An interior-point trust-region-based method for large-scale nonnegative regularization*, Inverse Problems, 18 (2002), pp. 1291–1307.

[26] H. Rutishauser, Lectures on Numerical Mathematics, M. Gutknecht, ed., Birkhäuser, Basel, 1990.

[27] Y. Saad, Iterative Methods for Sparse Linear Systems, 2nd ed., SIAM, Philadelphia, 2003.

[28] M. Stammberger and H. Voss, *On an unsymmetric eigenvalue problem governing free vibrations of fluid-solid structures*, Electron. Trans. Numer. Anal., 36 (2010), pp. 113–125.

[29] H. Voss, *An Arnoldi method for nonlinear eigenvalue problems*, BIT, 44 (2004), pp. 387–401.

[30] H. Wendland, Scattered Data Approximation, Cambridge University Press, Cambridge, 2005.