SIAM J. SCI. COMPUT.
Vol. 18, No. 4, pp. 1223–1241, July 1997

016

REGULARIZATION BY TRUNCATED TOTAL LEAST SQUARES*

R. D. FIERRO†, G. H. GOLUB‡, P. C. HANSEN§, AND D. P. O'LEARY¶

Abstract. The total least squares (TLS) method is a successful method for noise reduction in linear least squares problems in a number of applications. The TLS method is suited to problems in which both the coefficient matrix and the right-hand side are not precisely known. This paper focuses on the use of TLS for solving problems with very ill-conditioned coefficient matrices whose singular values decay gradually (so-called discrete ill-posed problems), where some regularization is necessary to stabilize the computed solution. We filter the solution by truncating the small singular values of the TLS matrix. We express our results in terms of the singular value decomposition (SVD) of the coefficient matrix rather than the augmented matrix. This leads to insight into the filtering properties of the truncated TLS method as compared to regularized least squares solutions. In addition, we propose and test an iterative algorithm based on Lanczos bidiagonalization for computing truncated TLS solutions.

Key words. total least squares, discrete ill-posed problems, regularization, bidiagonalization

AMS subject classifications. 65F20, 65F30

PII. S1064827594263837

1. Introduction.

The TLS method is a technique for solving overdetermined linear systems of equations. It was independently derived in several bodies of work, and is known by statisticians as the *errors in variables model*. Numerical analysts came to know it through the work of Golub and Van Loan [10] and Van Huffel and Vandewalle [24, 25, 26], and this literature has advanced the algorithmic and theoretical understanding of the method.

1.1. Motivation.

The development of the TLS technique was motivated by linear models $A\,x \approx b$ in which both the coefficient matrix A and the right-hand side b are subject to errors. In the TLS method one allows a residual matrix as well as a residual vector, and the computational problem becomes

$$\min_{\tilde{A},\tilde{b}} \|(A\,,\,b) - (\tilde{A}\,,\,\tilde{b})\|_F \qquad \text{subject to } \tilde{b} = \tilde{A}\,x\ . \tag{1}$$

Throughout the paper we assume that A is $m \times n$ with $m > n$. In contrast to the TLS formulation, the ordinary least squares (LS) method requires that $\tilde{A} = A$ and minimizes the 2-norm of the residual vector $b - \tilde{b}$.

*Received by the editors February 25, 1994; accepted for publication (in revised form) November 29, 1995.

http://www.siam.org/journals/sisc/18-4/26383.html

†Department of Mathematics, California State University, San Marcos, CA 92096 (fierro@thunder.csusm.edu). The work of this author was partially sponsored by a NATO collaborative research grant 5-2-05/RG900098.

‡Department of Computer Science, Stanford University, Stanford, CA 94305 (golub@sccm.stanford.edu). The work of this author was supported in part by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Air Force Office of Scientific Research under contract F49620–91–C–0086.

§Department of Mathematical Modelling, Building 305, Technical University of Denmark, DK-2800 Lyngby, Denmark (pch@imm.dtu.dk). The work of this author was partially sponsored by a NATO collaborative research grant 5-2-05/RG900098.

¶Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (oleary@cs.umd.edu).

The TLS technique has been traditionally applied to problems that are *numerically rank deficient*, i.e., where $(A\,,\,b)$ has one or more small (nonzero) singular values well separated from the large ones. The idea is to simply treat the small singular values of $(A\,,\,b)$ as zeros, reducing the problem to an exactly rank-deficient one [5, 26]. We shall call this technique *truncated TLS*. The technique is similar in spirit to truncated SVD, a natural generalization of the ordinary LS method for nearly rank-deficient problems that treats small singular values of A as zeros. In both methods, the almost redundant information in $(A\,,\,b)$ and A, respectively, associated with the small singular values, is discarded and the original ill-conditioned problem is replaced with another nearby and more well-conditioned problem with an exactly rank-deficient matrix. The major difference between the methods lies in the way that this is done: in truncated SVD the modification depends solely on A, while in truncated TLS the modification depends on both A and b.

Fierro and Bunch [5, 6] made a sensitivity analysis for the truncated TLS technique applied to a nearly rank-deficient A and showed how subspace sensitivity translates to solution sensitivity. The conclusion from their analysis is that truncated TLS is superior to truncated SVD when the right-hand side has large components corresponding to the small singular values that are retained (as in the full-rank case). An underlying assumption of this analysis is that the resulting rank-deficient system, obtained by deleting the small singular values, is well conditioned.

A related analysis which also focuses on the similarities between the truncated SVD and truncated TLS solutions of problems with well-defined numerical rank has been given by Wei [29, 30].

Many ill-conditioned problems arising in practical applications do not have a well-determined numerical rank; instead the singular values decay gradually to zero. Typically, these problems arise in connection with the numerical treatment of ill-posed problems, e.g., in spectroscopy, image processing, and nondestructive testing [12]. The discrete systems $A\,x \approx b$ derived from such ill-posed problems are often called discrete ill-posed problems, as they inherit many of the difficulties of the underlying ill-posed problem and therefore require a specialized treatment including some form of *regularization* [13] to suppress the effects of errors.

Most regularization methods used today assume that the errors are confined to the right-hand side b. Although this is true in many applications there are also problems in which the coefficient matrix A is not precisely known. For example, A may be available only by measurement or may be an idealized approximation of the true operator. Discretization typically also adds some errors to the matrix A. Hence, there is a need for developing methods that take into account the errors in A and their size relative to those in b. Since there is no gap in the singular value spectrum, the previous analyses have little to say about these problems, and the first goal of our work is to study properties of the truncated TLS solutions of discrete ill-posed problems. Our second goal is to develop practical computational algorithms for producing these solutions.

Thus, in this paper we investigate the truncated TLS technique and show that it produces a filtered solution. Moreover, we propose an iterative algorithm for computing the truncated TLS solution, based on Lanczos bidiagonalization. Our algorithm is efficient when the number of retained singular values of $(A\,,\,b)$ is small compared with the dimensions of A.

1.2. Stabilization by filtering. Most previous results about TLS have been stated in terms of the SVD of $(A\,,\,b)$, making comparisons with regularized least squares solutions difficult. The basis for our analysis in this paper is the SVD of A,

given by

$$A = U\,\Sigma\,V^T = \sum_{i=1}^{n} u_i\,\sigma_i\,v_i^T\ , \tag{2}$$

where $U = (u_1, \ldots, u_n)$ and $V = (v_1, \ldots, v_n)$ have orthonormal columns, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ with $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$, and the rank r of A is the number of strictly positive singular values.

The instabilities associated with discrete ill-posed problems can easily be illustrated. Consider the ordinary LS solution, which can be written as

$$x_{\mathrm{LS}} = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i}\,v_i\ .$$

Due to the division by small singular values σ_i, the solution x_{LS} may be dominated by components associated with the errors in b. Therefore, regularization is necessary to stabilize the solution. For example, in truncated SVD this is achieved by truncating the above sum at $k < n$:

$$x_k = \sum_{i=1}^{k} \frac{u_i^T b}{\sigma_i}\,v_i\ . \tag{3}$$

Tikhonov regularization [12, 13] is another well-known technique in which one solves the problem (with a given λ)

$$\min\ \left\{\ \|A\,x - b\|_2^2 + \lambda^2 \|L\,x\|_2^2\ \right\}\ , \tag{4}$$

where L is a matrix of full row rank used to control the size of the solution vector. It is easy to prove that if $L = I$, then the solution to (4) is given by

$$x_\lambda = \sum_{i=1}^{r} \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2}\,\frac{u_i^T b}{\sigma_i}\,v_i\ , \tag{5}$$

showing that this approach suppresses (filters) the components of the solution corresponding to the small singular values of A; see, e.g., [12, Section 5.1] or [15]. In this paper we prove that the same is true for truncated TLS.

Our paper is organized as follows. Section 2 summarizes the truncated TLS algorithm, and the filtering properties of this algorithm are analyzed in section 3. In section 4 we present an iterative algorithm based on Lanczos bidiagonalization that avoids the computation of the complete SVD of $(A\,,\,b)$. Regularization problems in general form are briefly discussed in section 5. Finally, in section 6 we present numerical results. We do not address the important issue of scaling of A and b (see instead [26, Section 3.6.2] for some details); neither do we address the choice of regularization parameter as it is beyond the scope of this paper.

2. Truncated TLS. We shall first explain what we mean by a truncated TLS solution, and in the next two sections we analyze this solution by means of the SVD.

The standard approach to TLS, developed by Golub and Van Loan [10], is based on the SVD of $(A\,,\,b)$. Recently, computationally cheaper techniques based on rank-revealing orthogonal decompositions have also appeared [2, 28]. For clarity, in this section we shall confine ourselves to the SVD-based approach and return to computational and algorithmic aspects in sections 4 and 5.

As mentioned in the Introduction, the approach taken in the truncated SVD technique is simply to neglect the small singular values of A. A similar approach is used in the truncated TLS method by neglecting the small singular values of $(A\,,\,b)$. To determine the number k of *large* singular values we can require a user-specified threshold or determine k adaptively; cf. section 4.2. The truncated TLS algorithm, given in [26, Section 3.6.1], can be summarized as follows.

ALGORITHM T-TLS.

1. Compute the SVD of the augmented matrix $(A\,,\,b)$:

$$(A\,,\,b) = \bar{U}\,\bar{\Sigma}\,\bar{V}^T = \sum_{i=1}^{n+1} \bar{u}_i\,\bar{\sigma}_i\,\bar{v}_i^T \tag{6}$$

   with $\bar{\sigma}_1 \geq \cdots \geq \bar{\sigma}_{n+1}$.

2. Choose a truncation parameter $k \leq \min(n, \operatorname{rank}(A\,,\,b))$ such that

$$\bar{\sigma}_k > \bar{\sigma}_{k+1} \qquad \text{and} \qquad \bar{V}_{22} \equiv (\bar{v}_{n+1,k+1}, \ldots, \bar{v}_{n+1,n+1}) \neq 0\ . \tag{7}$$

3. Partition the matrix $\bar{V}$ such that (with $q = n - k + 1$)

$$\bar{V} = \begin{array}{c} \overset{k}{\longleftrightarrow}\ \ \overset{q}{\longleftrightarrow} \\ \begin{pmatrix} \bar{V}_{11} & \bar{V}_{12} \\ \bar{V}_{21} & \bar{V}_{22} \end{pmatrix} \begin{array}{c} \updownarrow\ \ n \\ \updownarrow\ \ 1 \end{array} \end{array}\ . \tag{8}$$

4. Compute the minimum-norm TLS solution $\bar{x}_k$ as

$$\bar{x}_k = -\bar{V}_{12}\,\bar{V}_{22}^{\dagger} = -\bar{V}_{12}\,\bar{V}_{22}^T\,\|\bar{V}_{22}\|_2^{-2}\ . \tag{9}$$

In (9), $\bar{V}_{22}^{\dagger} = \bar{V}_{22}^T\,\|\bar{V}_{22}\|_2^{-2}$ denotes the pseudoinverse of $\bar{V}_{22}$ which exists because $\|\bar{V}_{22}\|_2 \neq 0$. The norms of $\bar{x}_k$ and the corresponding TLS residual matrix are given by

$$\|\bar{x}_k\|_2 = \sqrt{\left\|\bar{V}_{22}\right\|_2^{-2} - 1} \tag{10}$$

and

$$\|(A\,,\,b) - (\tilde{A}\,,\,\tilde{b})\|_F = \sqrt{\bar{\sigma}_{k+1}^2 + \cdots + \bar{\sigma}_{n+1}^2}\ , \tag{11}$$

where $\tilde{A}$ and $\tilde{b}$ are defined in (1). We see that $\|\bar{x}_k\|_2$ increases with k while the residual norm decreases with k.

If the second condition in (7) is violated, then k corresponds to a *nongeneric* problem (cf. [26, Section 3.4] for a definition). A more difficult situation is when the TLS problem is near nongeneric, for then $\|\bar{V}_{22}\|_2$ can become arbitrarily small and thus the solution norm $\|\bar{x}_k\|_2$ can become arbitrarily large. For this reason, it may be convenient to require that $\|\bar{V}_{22}\|_2$ is greater than some specified threshold τ which then limits the solution norm.

3. Filtering properties of the truncated TLS solution. In this section we take a closer look at the truncated TLS solution $\bar{x}_k$ and show how the SVD components corresponding to the small singular values are filtered. We will assume that the TLS problem associated with $A\,x \approx b$ is not near nongeneric—otherwise $\|\bar{V}_{22}\|_2$ can be very small and $\|\bar{x}_k\|_2$ therefore very large.

From [6, Theorem 4.1] we learn that

$$\|x_k - \bar{x}_k\|_2 \leq \mathcal{O}\left(\left(\frac{\bar{\sigma}_{k+1}}{\sigma_k}\right)^2\right) \sqrt{1+\|x_k\|^2}\,\sqrt{1+\|\bar{x}_k\|^2}\ .$$

If $k < n$ and there is a well-defined gap between σ_k and σ_{k+1} of A and if $\bar{\sigma}_{k+1}$ is close to σ_{k+1} (e.g., if the system $A\,x \approx b$ is almost consistent), then $\bar{\sigma}_{k+1}/\sigma_k \ll 1$ and hence the truncated TLS and the truncated SVD solutions are guaranteed to be similar (provided that the problem is not near nongeneric). Since x_k is a filtered solution we conclude that $\bar{x}_k$ is also a filtered solution.

For the discrete ill-posed problems that we are interested in, there is no gap in the singular value spectrum, and therefore we must use a different approach in our analysis of the filtering properties of $\bar{x}_k$. In order not to clutter our presentation with technical details and studies of special cases, we will assume that the nonzero singular values of A and $(A\,,\,b)$ are simple, i.e., $\sigma_1 > \sigma_2 > \cdots > \sigma_r > 0$, and similarly for the $\bar{\sigma}_i$. In this way, we can concentrate on the insight that we get from the derived results.

3.1. Technicalities. Our goal is to relate the truncated TLS solution $\bar{x}_k$ to the SVD of A by writing $\bar{x}_k$ as a filtered sum

$$\bar{x}_k = \sum_{i=1}^{r} f_i\,\frac{u_i^T b}{\sigma_i}\,v_i\ , \tag{12}$$

where f_i are the filter factors for truncated TLS. In order for this relation to be meaningful, we must show that $\bar{x}_k$ has no components along the vectors v_i corresponding to either $u_i^T b = 0$ or $i > r$. In the rest of this subsection, we prove this in a series of technical lemmas.

LEMMA 3.1. *If the nonzero singular values of A are simple, then a nonzero singular value of $(A\,,\,b)$ is equal to a singular value of A, i.e., $\bar{\sigma}_j = \sigma_i \neq 0$, if and only if $u_i^T b = 0$.*

Proof. First, we write $(A\,,\,b)^T(A\,,\,b)$ in the form

$$(A\,,\,b)^T(A\,,\,b) = \begin{pmatrix} V & 0 \\ 0 & 1 \end{pmatrix} \bar{\Lambda} \begin{pmatrix} V & 0 \\ 0 & 1 \end{pmatrix}^T$$

with $\bar{\Lambda}$ being a bordered diagonal matrix

$$\bar{\Lambda} = \begin{pmatrix} \Sigma^2 & \Sigma\,U^T b \\ b^T U\,\Sigma & \|b\|_2^2 \end{pmatrix}.$$

For $\sigma_i \neq 0$ we consider the matrix

$$\bar{\Lambda} - \sigma_i^2 I = \begin{pmatrix} \Sigma^2 - \sigma_i^2 I & \Sigma\,U^T b \\ b^T U\,\Sigma & \|b\|_2^2 - \sigma_i^2 \end{pmatrix},$$

where the ith diagonal element of $\Sigma^2 - \sigma_i^2 I$ is zero. The matrix $\bar{\Lambda} - \sigma_i^2 I$ is singular if σ_i is a singular value of $(A\,,\,b)$. We now use Gaussian elimination to annihilate the last row of $\bar{\Lambda} - \sigma_i^2 I$, except for the ith and the last element, and we obtain

$$\begin{pmatrix} \Sigma^2 - \sigma_i^2 I & \Sigma\,U^T b \\ \sigma_i u_i^T b e_{i,n}^T & \gamma \end{pmatrix},$$

where $e_{i,n}$ is the ith unit vector of dimension n, and γ is given by

$$\gamma = \|b\|_2^2 - \sigma_i^2 - \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\sigma_j^2 (u_j^T b)^2}{\sigma_j^2 - \sigma_i^2} \ .$$

We now interchange the ith and the last rows to obtain

$$\begin{pmatrix} \Sigma^2_{1:i-1,1:i-1} - \sigma_i^2 I & 0 & 0 & \Sigma_{1:i-1,1:i-1} U^T_{:,1:i-1} b \\ 0 & \sigma_i u_i^T b & 0 & \gamma \\ 0 & 0 & \Sigma^2_{i+1:n,i+1:n} - \sigma_i^2 I & \Sigma_{i+1:n,i+1:n} U^T_{:,i+1:n} b \\ 0 & 0 & 0 & \sigma_i u_i^T b \end{pmatrix}.$$

Obviously, if $\sigma_i \neq 0$, then this matrix is singular (i.e., $\sigma_i = \bar{\sigma}_j$ for some j) if and only if $u_i^T b = 0$. ∎

LEMMA 3.2. *If the nonzero singular values of both A and $(A\,,\,b)$ are simple, then $u_i^T b = 0 \Leftrightarrow \sigma_i = \bar{\sigma}_j \Leftrightarrow \bar{v}_j^T = (v_i^T, 0)$.*

Proof. From [26, Theorem 3.11] it follows that if the nonzero singular values of $(A\,,\,b)$ are simple and if $\sigma_i = \bar{\sigma}_j \neq 0$, then $u_i^T b = 0 \Leftrightarrow \bar{v}_j^T = (v_i^T, 0)$. Moreover, from Lemma 3.1 we know that if the nonzero singular values of A are simple, then $\sigma_i = \bar{\sigma}_j \neq 0 \Leftrightarrow u_i^T b = 0$. These two results lead to the result in Lemma 3.2. ∎

LEMMA 3.3. *If the nonzero singular values of A and $(A\,,\,b)$ are simple, and if $\sigma_i \neq 0$, then $u_i^T b = 0$ implies that $v_i^T \bar{x}_k = 0$.*

Proof. Let $\sigma_i = \bar{\sigma}_j \neq 0$ define the relation between i and j that we use throughout this proof. From Lemma 3.2 we have $\bar{v}_j^T = (v_i^T, 0)$, and from the orthogonality of $\bar{V}$ it then follows that

$$\bar{v}_j^T \begin{pmatrix} \bar{V}_{12} \\ \bar{V}_{22} \end{pmatrix} = v_i^T \bar{V}_{12} = \begin{cases} 0\,, & \text{if } j \leq k, \\ e^T_{j-k,n+1-k}\,, & \text{if } j > k\,, \end{cases}$$

where $e^T_{j-k,n+1-k}$ is the $(j-k)$th unit vector of dimension $n+1-k$. If $j \leq k$, we therefore have $v_i^T \bar{V}_{12}\, \bar{V}_{22}^T = 0$, and if $j > k$, we have $v_i^T \bar{V}_{12}\, \bar{V}_{22}^T = e^T_{j-k,n+1-k} \bar{V}_{22}^T = \bar{v}_{n+1,j} = 0$. Hence, we get $v_i^T \bar{x}_k = -v_i^T \bar{V}_{12}\, \bar{V}_{22}\, \|\bar{V}_{22}\|_2^{-2} = 0$, and since $\sigma_i = \bar{\sigma}_j \Leftrightarrow u_i^T b = 0$ we have proved that $u_i^T b = 0 \Rightarrow v_i^T \bar{x}_k = 0$. ∎

LEMMA 3.4. *If $\sigma_i = 0$, then $v_i^T \bar{x}_k = 0$.*

Proof. If $\sigma_i = 0$, then obviously $(A\,,\,b)$ also has a zero singular value with corresponding singular vector $(v_i^T, 0)^T$. Hence, following the proof for Lemma 3.3, we have that $v_i^T \bar{x}_k = 0$. ∎

LEMMA 3.5. *If $\sigma_i \neq 0$ and $u_i^T b \neq 0$, then $\bar{v}_{n+1,j} \neq 0$ implies that $\bar{\sigma}_j \neq \sigma_i$.*

Proof. We begin with the eigenequation $(A\,,\,b)^T (A\,,\,b) \bar{v}_j = \bar{\sigma}_j^2 \bar{v}_j$, which implies $(A^T A - \bar{\sigma}_j^2 I) \bar{v}_{1:n,j} = -\bar{v}_{n+1,j} A^T b$ and

$$(\Sigma^2 - \bar{\sigma}_j^2 I)\, V^T \bar{v}_{1:n,j} = -\bar{v}_{n+1,j} \Sigma\, U^T b$$

which is equivalent to the system of equations

$$(\sigma_i^2 - \bar{\sigma}_j^2)\, v_i^T \bar{v}_{1:n,j} = -\bar{v}_{n+1,j}\, \sigma_i\, u_i^T b\ , \qquad i = 1, \ldots, n\ .$$

If we assume that $\sigma_i \neq 0$ and $u_i^T b \neq 0$, then for any j we see that $\bar{v}_{n+1,j} \neq 0$ implies that the left-hand side is different from zero, and therefore $\bar{\sigma}_j \neq \sigma_i$. ∎

We have thus proved that (12) is a valid expression for $\bar{x}_k$. Moreover, we have shown that a singular value σ_i of A coincides with a singular value of $(A\,,\,b)$ if and only if $u_i^T b = 0$, and that $\bar{v}_{n+1,j} \neq 0$ implies that $\bar{\sigma}_j$ is not a singular value of A.

3.2. Main result: Filter factors. We can now state our main results about the filter factors f_i for truncated TLS.

THEOREM 3.6. *Let* (2) *be the SVD of the coefficient matrix* A *and* (6) *be the SVD of* $(A\,,\,b)$*, and suppose that the nonzero singular values of* A *and* $(A\,,\,b)$ *are simple. Then the filter factors* f_i *for* $\bar{x}_k$ *corresponding to* $u_i^T b \neq 0$ *and* $\sigma_i \neq 0$ *are given by*

$$f_i = \sum_{j=k+1}^{n+1} \frac{\bar{v}_{n+1,j}^2}{\left\|\bar{V}_{22}\right\|_2^2} \frac{\sigma_i^2}{\sigma_i^2 - \bar{\sigma}_j^2} \tag{13}$$

$$= \sum_{j=1}^{k} \frac{\bar{v}_{n+1,j}^2}{\left\|\bar{V}_{22}\right\|_2^2} \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} \ . \tag{14}$$

Proof. We see from (9) that the only columns $\bar{v}_j$ that contribute to the solution $\bar{x}_k$ are those for which $\bar{v}_{n+1,j} \neq 0$, i.e.,

$$\bar{x}_k = - \sum_{\substack{j=k+1 \\ \bar{v}_{n+1,j} \neq 0}}^{n+1} \frac{\bar{v}_{n+1,j}}{\left\|\bar{V}_{22}\right\|_2^2} \bar{v}_{1:n,j} \ .$$

Moreover, due to our assumptions and Lemma 3.5, the corresponding $\bar{\sigma}_j$ satisfy $\bar{\sigma}_j \neq \sigma_i$ for all i. By considering $(A\,,\,b)$ as an update of A, and using the secular equations from [1, Section 4.2], these columns can be written

$$\bar{v}_j = \frac{\bar{w}_j}{\|\bar{w}_j\|_2} \qquad \text{with } \bar{w}_j = \begin{pmatrix} \bar{w}_{1:n,j} \\ -1 \end{pmatrix}$$

and with $\bar{w}_{1:n,j}$ given by

$$\bar{w}_{1:n,j} = V\,(\Sigma^2 - \bar{\sigma}_j^2 I)^{-1} \Sigma\, U^T b = \sum_{i=1}^{n} \frac{\sigma_i\,(u_i^T b)}{\sigma_i^2 - \bar{\sigma}_j^2}\, v_i = \sum_{i=1}^{r} \frac{\sigma_i^2}{\sigma_i^2 - \bar{\sigma}_j^2} \frac{u_i^T b}{\sigma_i}\, v_i \ .$$

Using this relation and the fact that $\bar{v}_{n+1,j} = -\|\bar{w}\|_2^{-1}$, the expression for $\bar{x}_k$ then takes the form

$$\begin{aligned} \bar{x}_k &= - \sum_{\substack{j=k+1 \\ \bar{v}_{n+1,j} \neq 0}}^{n+1} \left(\frac{\bar{v}_{n+1,j}}{\left\|\bar{V}_{22}\right\|_2^2\, \|\bar{w}_j\|_2} \sum_{i=1}^{r} \frac{\sigma_i^2}{\sigma_i^2 - \bar{\sigma}_j^2} \frac{u_i^T b}{\sigma_i}\, v_i \right) \\ &= \sum_{i=1}^{r} \left(\sum_{\substack{j=k+1 \\ \bar{v}_{n+1,j} \neq 0}}^{n+1} \frac{\bar{v}_{n+1,j}^2}{\left\|\bar{V}_{22}\right\|_2^2} \frac{\sigma_i^2}{\sigma_i^2 - \bar{\sigma}_j^2} \right) \frac{u_i^T b}{\sigma_i}\, v_i \ . \end{aligned}$$

The expression in parentheses is the ith filter factor f_i in (12). From Lemma 3.5 we know that if $u_i^T b \neq 0$ and $i \leq r$, then $\bar{\sigma}_j \neq \sigma_i$ and therefore we can remove the requirement $\bar{v}_{n+1,j} \neq 0$ in the j-summation without worrying about dividing by zero. Hence, we have proved (13).

The proof for (14) is based on the secular equations associated with downdating the SVD of $(A\,,\,b)$ when b is deleted [1, Section 5]. For all the values of i that we consider, we know that $\bar{\sigma}_j \neq \sigma_i$, and the corresponding secular equations are

$$1 - \sum_{j=1}^{n+1} \frac{(\bar{u}_j^T b)^2}{\bar{\sigma}_j^2 - \sigma_i^2} = 0 \ .$$

From the relation $\bar{U}^T(A\,,\,b) = \bar{\Sigma}\,\bar{V}^T$ it follows immediately that $\bar{u}_j^T b = \bar{\sigma}_j \bar{v}_{n+1,j}$ for $j = 1, \ldots, n+1$. Hence, the secular equations become

$$1 - \sum_{j=1}^{n+1} \bar{v}_{n+1,j}^2 \, \frac{\bar{\sigma}_j^2}{\bar{\sigma}_j^2 - \sigma_i^2} = 0 \ .$$

Since $\bar{\sigma}_j^2/(\bar{\sigma}_j^2 - \sigma_i^2) = 1 + \sigma_i^2/(\bar{\sigma}_j^2 - \sigma_i^2)$, and since the $\bar{v}_{n+1,j}^2$ sum to one, this becomes

$$\sum_{j=1}^{n+1} \bar{v}_{n+1,j}^2 \, \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} = 0 \ .$$

Using this relation and rewriting (13) for the filter factors, we obtain

$$\begin{aligned} f_i &= \left\|\bar{V}_{22}\right\|_2^{-2} \sum_{j=1}^{n+1} \bar{v}_{n+1,j}^2 \frac{\sigma_i^2}{\sigma_i^2 - \bar{\sigma}_j^2} - \left\|\bar{V}_{22}\right\|_2^{-2} \sum_{j=1}^{k} \bar{v}_{n+1,j}^2 \frac{\sigma_i^2}{\sigma_i^2 - \bar{\sigma}_j^2} \\ &= \left\|\bar{V}_{22}\right\|_2^{-2} \sum_{j=1}^{k} \bar{v}_{n+1,j}^2 \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} \ . \end{aligned}$$

This is (14). ∎

Remark. If $k = n$ or $\bar{\sigma}_{k+1} = \cdots = \bar{\sigma}_{n+1}$, then (13) reduces to $f_i = \sigma_i^2/(\sigma_i^2 - \bar{\sigma}_{n+1}^2)$, consistent with [26, Theorem 2.7].

3.3. Bounds for the filter factors. We shall now give a further characterization of the filter factors for truncated TLS and thus show that $\bar{x}_k$ is indeed a filtered solution.

THEOREM 3.7. *Suppose that the nonzero singular values of A and $(A\,,\,b)$ are simple. Then the filter factors f_i for $i \le k$ corresponding to $u_i^T b \neq 0$ increase monotonically with i and satisfy*

$$0 \le f_i - 1 \le \frac{\bar{\sigma}_{k+1}^2}{\sigma_i^2 - \bar{\sigma}_{k+1}^2} \ . \tag{15}$$

Moreover, the filter factors f_i for $k < i \le r$ corresponding to $u_i^T b \neq 0$ satisfy

$$0 \le f_i \le \left\|\bar{V}_{22}\right\|_2^{-2} \frac{\sigma_i^2}{\bar{\sigma}_k^2 - \sigma_i^2} \ . \tag{16}$$

Proof. To prove (15) we first rewrite (13) in the form

$$f_i = \sum_{j=k+1}^{n+1} \frac{\bar{v}_{n+1,j}^2}{\left\|\bar{V}_{22}\right\|_2^2} + \sum_{j=k+1}^{n+1} \frac{\bar{v}_{n+1,j}^2}{\left\|\bar{V}_{22}\right\|_2^2} \left(\frac{\bar{\sigma}_j^2}{\sigma_i^2 - \bar{\sigma}_j^2} \right) = 1 + \sum_{j=k+1}^{n+1} \frac{\bar{v}_{n+1,j}^2}{\left\|\bar{V}_{22}\right\|_2^2} \left(\frac{\bar{\sigma}_j^2}{\sigma_i^2 - \bar{\sigma}_j^2} \right) \ .$$

It follows from the interlacing inequalities for the singular values of A and $(A\,,\,b)$ that $\sigma_i \ge \bar{\sigma}_{k+1}$ for $i = 1, \ldots, k$, and the inequality is sharp due to Lemma 3.2. Hence, the second term in the above equation is positive and we have proved the left inequality in (15). We also see that the filter factors increase with i because the singular values σ_i decrease with i. The right inequality follows from $\sum_{j=k+1}^{n+1} \bar{v}_{n+1,j}^2 = \left\|\bar{V}_{22}\right\|_2^2$ and the relation $\bar{\sigma}_j^2/(\sigma_i^2 - \bar{\sigma}_j^2) \le \bar{\sigma}_{k+1}^2/(\sigma_i^2 - \bar{\sigma}_{k+1}^2)$ for $j = k+1, \ldots, n+1$.

The proof for (16) is based on (14). With our assumptions we have $\sigma_k \neq \bar{\sigma}_{k+1}$, thus ensuring that f_i is positive for $i > k$. Inserting the relation $\sum_{j=1}^{k} \bar{v}_{n+1,j}^2 = \left\|\bar{V}_{21}\right\|_2^2 \leq 1$ into (14), we obtain

$$f_i \leq \left\|\bar{V}_{22}\right\|_2^{-2} \frac{\sigma_i^2}{\bar{\sigma}_k^2 - \sigma_i^2} \sum_{j=1}^{k} \bar{v}_{n+1,j}^2 \leq \left\|\bar{V}_{22}\right\|_2^{-2} \frac{\sigma_i^2}{\bar{\sigma}_k^2 - \sigma_i^2} \ .$$

Thus, we have proved (16). ∎

COROLLARY 3.8. *The norms of $\bar{x}_k$ and x_k satisfy*

$$\|\bar{x}_k\|_2 \geq \|x_k\|_2 \ , \qquad k = 1, \ldots, n \ . \tag{17}$$

Proof. Equation (17) is an immediate consequence of the fact that $f_i \geq 1$ for $i = 1, \ldots, k$, and $f_i \geq 0$ for $i = k+1, \ldots, n$. The corresponding filter factors for x_k are one and zero. ∎

From Theorem 3.7 we obtain the following expression for the first k filter factors:

$$1 \leq f_i \leq 1 + \frac{\bar{\sigma}_{k+1}^2}{\sigma_i^2} + \mathcal{O}\left(\frac{\bar{\sigma}_{k+1}^4}{\sigma_i^4}\right) \ , \quad i = 1, \ldots, k \ ,$$

showing that the larger the ratio between σ_i and $\bar{\sigma}_{k+1}$, the closer the bound on f_i is to one. Similarly, for the last $n-k$ filter factors we obtain

$$0 \leq f_i \leq \|\bar{V}_{22}\|_2^{-2} \frac{\sigma_i^2}{\bar{\sigma}_k^2} \left(1 + \mathcal{O}\left(\frac{\sigma_i^2}{\bar{\sigma}_k^2}\right)\right), \quad i = k+1, \ldots, n \ ,$$

showing that the smaller the ratio between σ_i and $\bar{\sigma}_k$, the closer f_i is to zero. Hence, Theorem 3.7 guarantees that the first k filter factors will be close to one and that the last $n-k$ filter factors will be small, even in the case where there is no large gap in the singular value spectrum, provided that $\|\bar{V}_{22}\|_2$ is not very small.

Thus, we have shown that $\bar{x}_k$ is a filtered solution because the contributions to $\bar{x}_k$ corresponding to all the small σ_i are filtered out while the remaining, significant contributions are retained in $\bar{x}_k$.

If $k = n$ and the errors in A and b are small, then the difference $\|x_{\mathrm{LS}} - \bar{x}_n\|_2$ between the LS and the TLS solutions is small [23], and our experience is that the same is true for $\|x_k - \bar{x}_k\|_2$ when $k < n$. However, when the noise is large, then $\bar{x}_k$ can be very different from x_k, and the filter factors f_i for $i \leq k$—especially f_k—can differ considerably from one (we have observed $f_k \approx 1.2$ in our experiments).

4. A bidiagonalization algorithm for large-scale problems. When the dimensions of A are not too large, one can compute the complete SVD of $(A\,,\,b)$ and then experiment with various choices of k. This is particularly useful if no a priori estimate of a suitable k is known.

When the dimensions of A become large, this approach becomes prohibitive because the SVD algorithm is of complexity $\mathcal{O}(mn^2)$. We shall therefore describe an alternative technique that is much better suited for large-scale problems whenever $k \ll n$, which is indeed the case in most discrete ill-posed problems.

A fairly straightforward approach would be to choose a sufficiently large $k_{\max}$ and compute a *partial SVD* of $(A\,,\,b)$, namely, the first $k_{\max}$ singular triplets $(\bar{\sigma}_i, \bar{u}_i, \bar{v}_i)$ of $(A\,,\,b)$. Then $\bar{x}_k$ can be computed by the alternative formula

$$\bar{x}_k = (\bar{V}_{11}^T)^{\dagger} \bar{V}_{21}^T \ . \tag{18}$$

The partial SVD can be computed by a technique similar to the PSVD algorithm described in [27] for computing the last few singular triplets. However, for large sparse or structured matrices (e.g., Toeplitz matrices, which arise in connection with discretization of many convolution problems) the partial SVD approach is prohibitive because this algorithm initially performs a reduction of $(A\,,\,b)$ to bidiagonal form, and the sparsity or structure of the matrix is lost in the first step of this reduction.

4.1. The Lanczos T-TLS algorithm. The above considerations lead us to consider iterative methods, based on Lanczos bidiagonalization, that do not alter the matrix A. It is well known that Lanczos bidiagonalization can be used to compute good approximations to the singular triplets associated with the largest singular values of a matrix; see, e.g., [9, 20]. We refer to the original papers and omit a discussion of the Lanczos bidiagonalization algorithm here. Again, we could choose some integer $k_{\max}$ and perform $k_{\max}$ Lanczos iterations applied to the augmented matrix $(A\,,\,b)$, after which we could compute approximate truncated TLS solutions for various k less than $k_{\max}$ by means of (18).

Here we propose an alternative technique based on Lanczos bidiagonalization of the matrix A rather than $(A\,,\,b)$. The key to our algorithm is to recognize that after k iterations, the Lanczos process with starting vector $u_1 = b/\|b\|_2$ has produced two sets of vectors $U_k = (u_1, \ldots, u_{k+1})$ and $V_k = (v_1, \ldots, v_k)$ and a $(k+1) \times k$ bidiagonal matrix B_k such that

$$A\,V_k = U_k B_k \qquad \text{and} \qquad \beta_1 u_1 = b\ .$$

Thus, after k Lanczos iterations we can project the TLS problem onto the subspaces spanned by U_k and V_k, in the hope that for large enough k we have captured all the large singular values of A that are needed for computing a useful regularized solution. The projected TLS problem is equivalent to

$$\min \left\| U_k^T((A\,,\,b) - (\hat{A}_k\,,\,\hat{b}_k)) \begin{pmatrix} V_k & 0 \\ 0 & 1 \end{pmatrix} \right\|_F \quad \text{subject to } U_k^T \hat{A}_k\, V_k\, y = U_k^T \hat{b}_k\ ,$$

or

$$\min \|(B_k\,,\,\beta_1 e_1) - (\hat{B}_k\,,\,\hat{e}_k)\|_F \qquad \text{subject to } \hat{B}_k\, y = \hat{e}_k\ , \tag{19}$$

where $e_1 = (1, 0, \ldots, 0)^T$, and $\hat{B}_k$ and $\hat{e}_k$ are generally full. Our algorithm reduces to the LSQR algorithm [21] if we require $\hat{B}_k = B_k$ in each step.

In each Lanczos step we can now compute an approximate truncated TLS solution $\tilde{x}_k$ by applying the Algorithm TLS to the small-size problem in (19). Hence, we compute the SVD of the matrix $(B_k\,,\,\beta_1 e_1)$,

$$(B_k\,,\,\beta_1 e_1) = \bar{\bar{U}}^{(k)}\, \bar{\bar{\Sigma}}^{(k)} \left(\bar{\bar{V}}^{(k)}\right)^T\ , \qquad \bar{\bar{V}}^{(k)} = \begin{array}{c} \\ \\ \end{array} \begin{array}{c} \overset{k}{\longleftrightarrow} \\ \\ \end{array} \begin{array}{c} \overset{1}{\longleftrightarrow} \\ \\ \end{array} \begin{pmatrix} \bar{\bar{V}}_{11}^{(k)} & \bar{\bar{V}}_{12}^{(k)} \\ \bar{\bar{V}}_{21}^{(k)} & \bar{\bar{v}}_{22}^{(k)} \end{pmatrix} \begin{array}{c} \updownarrow \\ \updownarrow \end{array} \begin{array}{c} k \\ 1 \end{array}\ ,$$

and the standard TLS solution $\bar{y}_k$ to (19) is

$$\bar{y}_k = -\bar{\bar{V}}_{12}^{(k)} \left(\bar{\bar{v}}_{22}^{(k)}\right)^{-1}\ .$$

Then the approximate TLS solution $\tilde{x}_k$ is given by

$$\tilde{x}_k = -V_k\, \bar{y}_k = -V_k\, \bar{\bar{V}}_{12}^{(k)} \left(\bar{\bar{v}}_{22}^{(k)}\right)^{-1}\ . \tag{20}$$

For convenience, we can permute the vector $\beta_1 e_1$ in front of B_k such that, in each step, we merely need to compute the last singular triplet of the $(k+1)\times(k+1)$ upper bidiagonal matrix $(\beta_1 e_1\,,\,B_k)$. This can be done in $\mathcal{O}(k^2)$ operations by means of the PSVD algorithm [27].

We remark that it is easy to augment the above algorithm to include the computations of the LSQR algorithm [21]. Approximate truncated SVD solutions can be computed together with the approximate T-TLS solutions with little extra overhead.

4.2. Stopping criterion. During the iterations it is helpful to display the norms of the solution vector $\tilde{x}_k$ and the corresponding TLS residual matrix. Both norms are easy to express in terms of the SVD of $(B_k\,,\,\beta_1 e_1)$ and require very little computational effort.

THEOREM 4.1. *The norms of the solution and the residual matrix in the Lanczos T-TLS algorithm satisfy*

$$\|\tilde{x}_k\|_2 = \sqrt{\left(\bar{\bar{v}}_{22}^{(k)}\right)^{-2} - 1} \tag{21}$$

and

$$\|(A\,,\,b) - (\hat{A}_k\,,\,\hat{b}_k)\|_F^2 = \|(A\,,\,b)\|_F^2 - \|(B_k\,,\,\beta_1 e_1)\|_F^2 + (\bar{\bar{\sigma}}_{k+1}^{(k)})^2\ , \tag{22}$$

where $\bar{\bar{\sigma}}_{k+1}^{(k)}$ *is the smallest singular value of* $(B_k\,,\,\beta_1 e_1)$. *Moreover,* $\|\tilde{x}_k\|_2$ *is a nondecreasing function of* k *and the residual norm in* (22) *is a nonincreasing function of* k.

Proof. Equations (21) and (22) follow immediately from the SVD of $(B_k\,,\,\beta_1 e_1)$. That the residual norm cannot increase is an immediate consequence of the interlacing inequalities for the singular values of $(B_k\,,\,\beta_1 e_1)$ and $(B_{k+1}\,,\,\beta_1 e_1)$. To prove that $\|\tilde{x}_k\|_2$ cannot decrease with k we must show that $|\bar{\bar{v}}_{22}^{(k)}| \geq |\bar{\bar{v}}_{22}^{(k+1)}|$ for all k. This is proved in the Appendix. ∎

We remark that for the LSQR algorithm, the norm of the residual vector is monotonically decreasing, since we minimize over an expanding subspace [21]. Further, since LSQR is mathematically equivalent to applying the conjugate gradient method to the normal equations, the fact that the solution norm is monotonically increasing follows from equation (6:3) of Hestenes and Stiefel [19].

Notice that (22) is only guaranteed to hold in exact arithmetic while it fails to hold in inexact arithmetic when spurious singular values of $(A\,,\,b)$ start to appear in $(B_k\,,\,\beta_1 e_1)$. The cure is either to use selective reorthogonalization or to identify the spurious singular values; see the discussion in [3, Chapter 2].

The Lanczos iteration gives us a sequence of truncated TLS solutions $\{\tilde{x}_k\}$.[1] We need a criterion for choosing a good stopping index k. If explicit knowledge about the errors in A and b is available, then this information can be used to stop when the norm of the TLS residual matrix equals its expected value—similar to the so-called discrepancy principle for LS problems; see [13, Section 5.3]. Here we are concerned with the situation where no knowledge about the noise in A and b is available, so that this information, in a sense, has to be extracted from the given data.

A conceptually simple stopping criteria is to stop when the norm of the residual vector—in our case $\|A\,\tilde{x}_k - b\|_2$—is considered small, e.g., when it levels off at some

[1] At each iteration we could also truncate at singular value $\hat{k} < k$, producing a set of T-TLS solutions $\{\tilde{x}_{\hat{k},k}\}$ for $k = 1, 2, \ldots$ and $\hat{k} = 1, 2, \ldots, k$, but we do not pursue this idea here.

value reflecting the errors. This is a quite useful stopping rule for well-conditioned least squares problems because the solution vector for such problems changes slowly from step to step, and hence the precise choice of k is not so important. On the other hand, for discrete ill-posed problems this criterion is more likely to fail because the solution vector for such problems may change dramatically in each iteration step as the residual norm approaches its stalling phase. Nevertheless, we have actually had some success with this stopping rule; see section 6.

Another popular method for choosing the regularization parameter is the method of generalized cross-validation due to Golub, Heath, and Wahba [8]. Currently, we do not have any experience with this method when applied to our algorithms.

A third possible stopping criterion can be based on the L-curve criterion studied recently in [16, 18]. The idea in this method is to plot in log–log scale the solution norm versus the residual norm, in our case $\|\tilde{x}_k\|_2$ versus $\|(A\,,\,b) - (\hat{A}_k\,,\,\hat{b}_k)\|_F$, and choose as the optimal k the truncation parameter at which this curve has an L-shaped *corner*. Essentially, the corner is defined by locating the point with greatest curvature in the log–log scale. For more information on this technique, see [18].

Of course, the L-curve's corner cannot be identified without going a few steps too far, but we believe that any good stopping criterion for discrete ill-posed problems (including generalized cross-validation) will suffer from this mild inconvenience.

5. Regularization in general form. Theorems 3.6 and 3.7 show that the T-TLS solution $\bar{x}_k$ is a filtered solution whose main contributions come from the first k right singular vectors v_i. It is common knowledge that these vectors are not always the best basis vectors for a regularized solution. This is the reason for using a matrix $L \neq I$ in Tikhonov regularization (4), commonly called regularization in general form. Then it is convenient to introduce the quotient SVD (QSVD)[2] of the matrix pair (A, L):

$$A = \breve{U}\,\mathrm{diag}(\alpha_i)\,W^{-1}\,, \qquad L = \breve{V}\,\mathrm{diag}(\beta_i)\,W^{-1}\,, \tag{23}$$

for then the regularized solution is expanded in terms of the columns w_i of W, and the main contributions come from the vectors w_i associated with the largest generalized singular values α_i/β_i; see, e.g., [14], [13, Section 4], or [17, Section 6] for details.

In connection with our T-TLS algorithms it may also be convenient to implicitly use regularization in general form with $L \neq I$. This is done in the same way as general-form regularization is treated in connection with Tikhonov regularization and other methods. First, transform the problem involving A, L, and b into a standard-form problem with matrix A_{sf} and right-hand side b_{sf}. Then apply T-TLS or Lanczos T-TLS to the standard-form problem to obtain a regularized solution x_{sf}. Finally, transform x_{sf} back to the general-form setting.

There are several ways to transform a problem into standard form. The following transformation originally due to Eldén [4] is well suited. Let

$$L_A^{\dagger} = W\,\mathrm{diag}(\beta_i^{-1})\,\breve{V}^T$$

denote the A-weighted generalized inverse of L; cf. [4] for a formal definition. Then A_{sf} and b_{sf} are given by

$$A_{\mathrm{sf}} = A\,L_A^{\dagger} = \breve{U}\,\mathrm{diag}(\alpha_i\beta_i^{-1})\,\breve{V}^T\,, \qquad b_{\mathrm{sf}} = b - A\,x_0\,, \tag{24}$$

[2]The QSVD is also commonly referred to as the generalized SVD (GSVD).

where x_0 is the component of the solution in the null space of L (this vector can easily be computed a priori). Moreover, the transformation back to the general-form setting essentially requires a multiplication with $L_A^\dagger$:

$$x = L_A^\dagger x_{\mathrm{sf}} + x_0 \ . \tag{25}$$

When the T-TLS algorithm is applied to the standard-form problem, then

$$\bar{x}_{\mathrm{sf},k} = \sum_{i=1}^{\ell} f_{\mathrm{sf},i} \, \frac{\breve{u}_i^T b_{\mathrm{sf}}}{\alpha_i \beta_i^{-1}} \, \breve{v}_i \ ,$$

where ℓ is the row rank of L, and $f_{\mathrm{sf},i}$ are the filter factors associated with the application of T-TLS to $(A_{\mathrm{sf}}\,,\, b_{\mathrm{sf}})$. Moreover, we get

$$\bar{x}_k = \sum_{i=1}^{\ell} f_{\mathrm{sf},i} \, \frac{\breve{u}_i^T b_{\mathrm{sf}}}{\alpha_i} \, w_i + x_0 \ .$$

When L is well conditioned (which is the usual case in regularization problems), then the generalized singular values of (A, L) decay gradually to zero in the same manner as the singular values of A. Some insight into this phenomenon can be found in [14], and as a consequence the filter factors $f_{\mathrm{sf},i}$ essentially filter out the contributions to $\bar{x}_k$ corresponding to the small generalized singular values. Hence, $\bar{x}_k$ is indeed a general-form regularized solution.

The key to the efficiency of this method in connection with the Lanczos T-TLS algorithm is that the matrix A_{sf} is never formed explicitly; we only need to be able to perform matrix–vector multiplications with A, A^T, $L_A^\dagger$, and $(L_A^\dagger)^T$. Given a basis N for the null space of L, the latter two matrix multiplications can be done in $\mathcal{O}((n-\ell)n)$ operations, as long as L is a banded matrix, by means of the following algorithms:

COMPUTE $y = L_A^\dagger x$

1. $y \leftarrow \begin{pmatrix} I_{n-\ell} & 0 \\ & L \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ x \end{pmatrix}$
2. $y \leftarrow y - N\,T\,y$

COMPUTE $y = (L_A^\dagger)^T x$

1. $x \leftarrow x - T^T N^T x$
2. $\begin{pmatrix} y \\ z \end{pmatrix} \leftarrow \begin{pmatrix} L \\ 0 & I_{n-\ell} \end{pmatrix}^{-T} x$

where the $(n-\ell) \times n$ matrix $T = (AN)^\dagger A$ is computed only once in $\mathcal{O}(mn(n-\ell))$ operations. The work in the computation of x_0 is dominated by $n-\ell$ multiplications with A. We omit the details here and refer to the discussion of implementation details given in [13, Section 4.3].

6. Numerical examples. In this section we illustrate the use of the T-TLS and Lanczos T-TLS algorithms for solving discrete ill-posed problems. We compare the solutions computed by these two methods with the solutions from three classical methods for discrete ill-posed problems, namely, Tikhonov regularization, truncated SVD, and LSQR. Our experiments were carried out in MATLAB using the REGULARIZATION TOOLS package [17].

Our test problems were generated as follows. The matrix A is 64×32 and comes from discretization of Phillips's test problem (cf. [17, phillips]). Two right-hand sides $b^{[1]}$, $b^{[2]}$ were generated artificially by means of the SVD of A. The Fourier coefficients $\eta_i^{[1]} = u_i^T b^{[1]}$ of the first satisfy

$\eta_1^{[1]}, \ldots, \eta_8^{[1]}$ are geometrically spaced between 10^{-4} and 1,
$\eta_8^{[1]}, \ldots, \eta_{32}^{[1]}$ are geometrically spaced between 1 and 10^{-20}.

For the second,

$\eta_1^{[2]}, \ldots, \eta_{32}^{[2]}$ are geometrically spaced between 1 and 10^{-16}.

Here geometrically spaced means that the ratio $\eta_i^{[p]}/\eta_{i+1}^{[p]}$ is a constant. Only $b^{[1]}$ has coefficients $\eta_i^{[1]}$ that increase with i, and from the theory in [5, 6] we therefore expect that TLS is superior to LS for $b^{[1]}$ only. Both systems are scaled such that $\max_{ij} |a_{ij}| = \max_i |b_i^{[p]}| = 1$ and the corresponding exact solutions are $x_{\text{exact}}^{[p]} = A^\dagger b^{[p]}$. Then we add perturbations E and e with elements from a Gaussian distribution with zero mean and standard deviation chosen such that $\|E\|_2 = \|e\|_2 = \epsilon$, where ϵ is a specified constant.

When we perturb the matrix A randomly as described above, and if the noise level is large, then the singular vectors of the perturbed A are approximately equal to the corresponding unperturbed singular vector plus a high-frequency component that clearly resembles the Gaussian noise added to the unperturbed matrix. This follows from a Taylor expansion of the singular vectors with respect to the elements of the perturbation matrix E.

An important consequence of the above perturbation of the SVD is that standard-form regularization with $L = I$ is not suited because the high-frequency component appearing in all singular vectors also appears in the regularized solutions, no matter which regularization method is used and how the regularization parameter is chosen. The only way to avoid the high-frequency part in the regularized solutions is to use a different regularization matrix. We have chosen L equal to the approximate second derivative operator; i.e., L is $(n-2)\times n$ and has rows of the form $(\ldots, 0, 1, -2, 1, 0, \ldots)$. The transformation to and from standard form was carried out as explained in section 5 using the implementations gen_form and std_form from [17].

For each combination of ϵ and right-hand side b we generated 1000 test problems, and each test problem was solved by means of the following regularization methods:

1. T-TLS with $k = 1, \ldots, 12$.
2. Lanczos T-TLS with $k_{\max} = 12$ iterations and complete reorthogonalization.
3. Tikhonov regularization with λ in the range $(10^{-8}, 10^2)$.
4. Truncated SVD with $k = 1, \ldots, 12$.
5. The LSQR algorithm with $k_{\max} = 12$ iterations.

First, we want to compare the optimal accuracy that can be attained by any of the above methods. To do this, for each method we define the optimal regularized solution x_{opt} as the one closest to the exact solution. For example, for T-TLS,

$$\|\bar{x}_{\text{opt}} - x_{\text{exact}}\|_2 \le \|\bar{x}_k - x_{\text{exact}}\|_2 \ , \qquad k = 1, \ldots, 12 \ .$$

In this way, we can investigate under which circumstances the TLS approach is capable of outperforming the LS approach.

Test 1. This test was carried out with a relatively "large" noise level $\epsilon = 5 \cdot 10^{-2}$ and with the first right-hand side $b^{[1]}$ for which the first eight coefficients $\eta_i^{[1]}$ increase. Figure 1 shows histograms of the relative errors $\|x_{\text{opt}} - x_{\text{exact}}\|_2 / \|x_{\text{exact}}\|_2$ for all five regularization methods. It is evident that for this test problem, both the T-TLS and the Lanczos T-TLS algorithms are able to produce more accurate solutions than the three classical regularization methods. Moreover, we see that T-TLS and Lanczos T-TLS produce almost the same histograms—and the same is true for the other three methods.

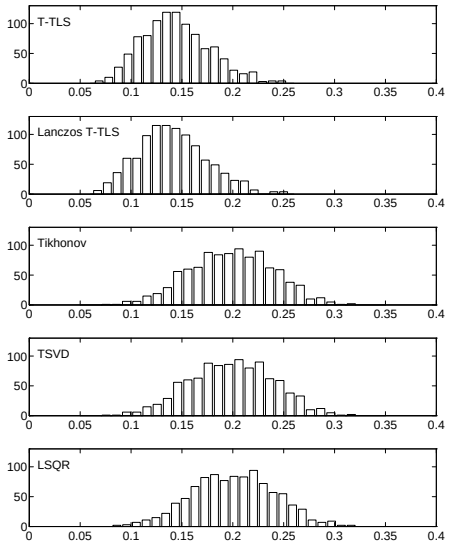

FIG. 1. *Test* 1: *error level* $\epsilon = 5 \cdot 10^{-2}$ *and right-hand side* $b^{[1]}$. *Histograms for the optimal relative errors of* 1000 *test problems solved by five different regularization methods. Algorithms T-TLS and Lanczos T-TLS are superior to the three classical methods.*

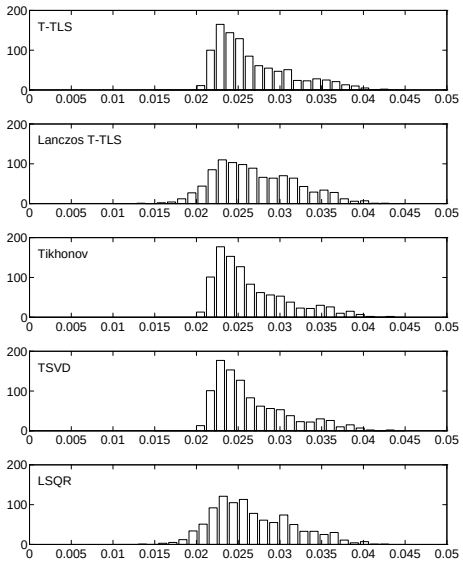

FIG. 2. *Test* 2: *error level* $\epsilon = 10^{-3}$ *and right-hand side* $b^{[1]}$. *Histograms for the optimal relative errors of* 1000 *test problems solved by five different regularization methods. All five methods give essentially the same results.*

Test 2. Our second test problem is identical to the first problem, except that the noise level is now smaller, $\epsilon = 10^{-3}$. It is well known that for small noise levels, we should not expect much difference in the TLS and LS solutions. The results in Fig. 2 confirm this: even though the right-hand side is the same as in Test 1, the histograms for all five methods are now almost identical. Notice, in particular, the resemblance

TABLE 1
Average flop counts for three test problems.

	Test 1	Test 2	Test 3
Full SVD	$7.8 \cdot 10^5$	$6.7 \cdot 10^5$	$7.9 \cdot 10^5$
Bidiagonalization	$4.0 \cdot 10^5$	$4.0 \cdot 10^5$	$4.0 \cdot 10^5$
Lanczos T-TLS	$0.9 \cdot 10^5$	$1.7 \cdot 10^5$	$0.5 \cdot 10^5$

of T-TLS, Tikhonov and truncated SVD, and the resemblance of Lanczos T-TLS and LSQR.

Test 3. Our final test problem uses the second right-hand side $b^{[2]}$ for which all the Fourier coefficients $u_i^T b^{[2]}$ decay, and the same "large" noise level as in Test 1. All five histograms (not shown here) are almost identical, illustrating that for this class of problems, we cannot expect the TLS approach to outperform the LS approach.

These examples illustrate that the TLS technique can indeed produce results that are superior to those computed by the classical regularization methods, when the noise is large and the right-hand has large SVD components corresponding to the smallest retained singular values. Moreover, we have seen that the iterative Lanczos T-TLS algorithm can produce results which are very similar to those obtained by the much more expensive T-TLS algorithm that requires a (partial) SVD computation.

We have also illustrated that when the right-hand side does not have large SVD components corresponding to the smallest retained singular values, or when the errors are "small," then there is no advantage in using the TLS approach over the classical methods.

To illustrate that the Lanczos T-TLS procedure can be much more efficient than the T-TLS procedure, Table 1 shows the average number of flops used in the Lanczos T-TLS procedure as well as in a full SVD computation and in a full bidiagonalization (excluding the standard-form transformation). The last two flop counts are lower bounds for the computational work involved in computing the T-TLS solution via a full and partial SVD, respectively. In the third test problem, the Lanczos T-TLS procedure is almost 10 times faster than the T-TLS procedure using a partial SVD. The ratio can easily be much bigger when the procedures are applied to sparse or structured matrices.

Next, we briefly report on our experience with choosing a good regularization parameter k for T-TLS and Lanczos T-TLS.

For Test 1, we found that plots of the solution norm versus the norm of the TLS residual matrix or the TLS residual vector do not have any L-shape, as required in the L-curve criterion. Instead, we obtained good results when stopping the iteration process when the norm of the residual vector, $\|A\,\tilde{x}_k - b\|_2$, levels off. In fact, in our experiments $\|A\,\tilde{x}_k - b\|_2$ always reached a minimum for some small value of k, after which it increased slowly again, and this minimum was used to choose k. When we compare the optimal errors with the errors obtained by using this simple parameter choice rule, we obtain essentially the same results and histograms (not shown here).

For Tests 2 and 3, we find that the L-curve criterion works well when we plot the norm of the solution versus the norm of the TLS residual matrix. We refer to [16, 18] for numerical examples. Further research in this area is required.

7. Conclusion. We have demonstrated that the T-TLS method has a filtering effect when applied to discrete ill-posed problems and that the method in some cases is superior to the truncated SVD method. We have also presented an iterative algorithm for computing approximate T-TLS solutions, based on Lanczos bidiagonalization,

which can be much more efficient than the SVD-based T-TLS algorithm for sparse and structured matrices.

Appendix. In this appendix we complete the proof of Theorem 4.1 by proving that $|\bar{\bar{v}}_{22}^{(k)}| \geq |\bar{\bar{v}}_{22}^{(k+1)}|$ for all $k > 0$.[3] We introduce the following notation:

$$T_k \equiv (\beta_1 e_1 \, , \, B_k)^T (\beta_1 e_1 \, , \, B_k) \; , \quad s_k \equiv \bar{\bar{\sigma}}_{k+1}^{(k)} \; , \quad s_{k+1} \equiv \bar{\bar{\sigma}}_{k+2}^{(k+1)} \; ,$$

and the first column of B_k is denoted $(\alpha_1, \beta_2, 0, \ldots)^T$. Then T_k is a tridiagonal symmetric positive definite $(k+1) \times (k+1)$ matrix with eigenvalues $(\bar{\bar{\sigma}}_1^{(k)})^2, \ldots, (\bar{\bar{\sigma}}_{k+1}^{(k)})^2$. Due to the Lanczos process all elements of B_k are nonnegative, and it follows that T_k is an oscillatory matrix [7, Chapter XIII, Section 9] and that the eigenvector w associated with the smallest eigenvalue s_k^2 has k sign changes [7, p. 105], i.e.,

$$\text{sign}(w_{i+1}) = -\text{sign}(w_i) \; , \quad i = 1, \ldots, k \; .$$

Moreover, we can always choose w such that $w_1 \geq 0$. The following two lemmas lead to the desired result.

LEMMA A.1. *Let τ_i denote the diagonal elements of T_k. Then*

$$s_k^2 \leq \min_i \tau_i \quad \textit{for} \quad k > 0 \; . \tag{A.26}$$

Proof. We know that $s_k^2 \leq z^T T_k z$ for any vector z of length one. Choosing z as the ith unit vector yields this familiar result. □

LEMMA A.2. *Fix k and let w and z be eigenvectors such that*

$$T_k w = s_k^2 w \qquad \textit{and} \qquad T_{k+1} z = s_{k+1}^2 z$$

with $\|w\|_2 = \|z\|_2 = 1$, $w_1 \geq 0$, and $z_1 \geq 0$. Then

$$w_1 - z_1 \geq 0 \; . \tag{A.27}$$

Proof. Our proof strategy will be to show that if we normalize so that $w_i = z_i$, then $|w_{i+1}| < |z_{i+1}|$. It then follows that renormalization to $\|w\|_2 = \|z\|_2 = 1$ yields (A.27).

Let $w_1 = z_1 = 1$. Denote the nonzeros in the ith row of T_k by $(\gamma_i, \tau_i, \gamma_{i+1})$. Then the first row yields the relations

$$\begin{aligned} \tau_1 w_1 + \gamma_2 w_2 &= s_k^2 w_1 \; , \\ \tau_1 z_1 + \gamma_2 z_2 &= s_{k+1}^2 z_1 \; , \end{aligned}$$

so

$$\begin{aligned} w_2 &= \frac{s_k^2 - \tau_1}{\gamma_2} \; , \\ z_2 &= \frac{s_{k+1}^2 - \tau_1}{\gamma_2}, \end{aligned}$$

[3] This result can also be established as a consequence of equation (3.4.8) in Szegö [22] by noting that $|\bar{\bar{v}}_{22}^{(k)}|$ and $|\bar{\bar{v}}_{22}^{(k+1)}|$ are the square roots of the Christoffel numbers λ_{1k} and $\lambda_{1,k+1}$ [11], but we prefer a direct matrix algebra proof.

so $z_2 < w_2 < 0$. A similar computation for the second row yields

$$w_3 = \frac{(\tau_2 - s_k^2)(-w_2) - \gamma_2}{\gamma_3} ,$$
$$z_3 = \frac{(\tau_2 - s_{k+1}^2)(-z_2) - \gamma_2}{\gamma_3} ,$$

and therefore $0 < w_3 < z_3$.

There is a stronger monotonicity relation

$$\begin{aligned} \frac{z_3}{z_2} - \frac{w_3}{w_2} &= \frac{1}{\gamma_3} \left\{ -(\tau_2 - s_{k+1}^2) - \frac{\gamma_2}{z_2} + (\tau_2 - s_k^2) + \frac{\gamma_2}{w_2} \right\} \\ &= \frac{1}{\gamma_3} \left\{ (s_{k+1}^2 - s_k^2) + \gamma_2 \left(\frac{1}{w_2} - \frac{1}{z_2} \right) \right\} \\ &< 0 , \end{aligned}$$

since both quantities in parentheses are negative.

This is the setup for an induction argument. Assume, for convenience, that we renormalize so that $w_i = z_i = 1$ $(i < k-1)$, and assume that the renormalized vector satisfies $z_{i+1} < w_{i+1} < 0$. Then the same argument, using the $(i+1)$st row of the matrix, yields $0 < w_{i+2} < z_{i+2}$ and

$$\frac{z_{i+2}}{z_{i+1}} - \frac{w_{i+2}}{w_{i+1}} < 0 ,$$

completing the induction. □

The result about $|\bar{\bar{v}}_{22}^{(k)}| \geq |\bar{\bar{v}}_{22}^{(k+1)}|$ now follows immediately by recognizing that the eigenvectors associated with $s_k^2 = (\bar{\bar{\sigma}}_{k+1}^{(k)})^2$ and $s_{k+1}^2 = (\bar{\bar{\sigma}}_{k+2}^{(k+1)})^2$ are

$$w = \begin{pmatrix} \bar{\bar{v}}_{22}^{(k)} \\ \bar{\bar{V}}_{12}^{(k)} \end{pmatrix} \qquad \text{and} \qquad z = \begin{pmatrix} \bar{\bar{v}}_{22}^{(k+1)} \\ \bar{\bar{V}}_{12}^{(k+1)} \end{pmatrix} ,$$

i.e., cyclic permutations of the last column of $\bar{\bar{V}}^{(k)}$ and $\bar{\bar{V}}^{(k+1)}$ from section 4.1. Thus, $|\bar{\bar{v}}_{22}^{(k)}| - |\bar{\bar{v}}_{22}^{(k+1)}| = w_1 - z_1 \geq 0$ for all $k > 0$.

Acknowledgments. We are grateful for helpful comments from Stan Eisenstat, Chris Paige, Zdenek Strakos, and the referees.

REFERENCES

[1] J. R. Bunch and C. P. Nielsen, *Updating the singular value decomposition*, Numer. Math., 31 (1978), pp. 111–129.

[2] T. F. Chan and P. C. Hansen, *Some applications of the rank revealing QR factorization*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 727–741.

[3] J. K. Cullum and R. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations; Vol.* I, *Theory*, Birkhäuser Boston, Cambridge, MA, 1985.

[4] L. Eldén, *A weighted pseudoinverse, generalized singular values, and constrained least squares problems*, BIT, 22 (1982), pp. 487–502.

[5] R. D. Fierro and J. R. Bunch, *Collinearity and total least squares*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1167–1181.

[6] R. D. Fierro and J. R. Bunch, *Perturbation theory for orthogonal projection methods with application to least squares and total least squares*, Linear Algebra Appl., 234 (1996), pp. 71–96.

[7] F. R. Gantmacher, *The Theory of Matrices; Vol.* II, Chelsea, New York, 1959.
[8] G. H. Golub, M. T. Heath, and G. Wahba, *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics, 21 (1979), pp. 215–223.
[9] G. H. Golub and W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, SIAM J. Numer. Anal., 2 (Series B) (1965), pp. 205–224.
[10] G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.
[11] G. Golub and J. Welsch, *Calculation of Gauss quadrature rules*, Math. Comp., 23 (1969), pp. 221–230.
[12] C. W. Groetsch, *Inverse Problems in the Mathematical Sciences*, Vieweg, Wiesbaden, 1993.
[13] M. Hanke and P. C. Hansen, *Regularization methods for large-scale problems*, Surveys Math. Indust., 3 (1993), pp. 253–315.
[14] P. C. Hansen, *Regularization, GSVD and truncated GSVD*, BIT, 29 (1989), pp. 491–504.
[15] P. C. Hansen, *Truncated SVD solutions to discrete ill-posed problems with ill-determined numerical rank*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 503–518.
[16] P. C. Hansen, *Analysis of discrete ill-posed problems by means of the L-curve*, SIAM Rev., 34 (1992), pp. 561–580.
[17] P. C. Hansen, *Regularization tools: A Matlab package for analysis and solution of discrete ill-posed problems*, Numer. Algorithms, 6 (1994), pp. 1–35.
[18] P. C. Hansen and D. P. O'Leary, *The use of the L-curve in the regularization of discrete ill-posed problems*, SIAM J. Sci. Comput., 14 (1993), pp. 1487–1503.
[19] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bureau Standards, 49 (1952), pp. 409–436.
[20] D. P. O'Leary and J. A. Simmons, *A bidiagonalization-regularization procedure for large scale discretizations of ill-posed problems*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 474–489.
[21] C. C. Paige and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71.
[22] G. Szegö, *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1939.
[23] G. W. Stewart, *On the invariance of perturbed null vectors under column scaling*, Numer. Math., 44 (1984), pp. 61–65.
[24] S. Van Huffel and J. Vandewalle, *Algebraic relationships between classical regression and total least-squares estimation*, Linear Algebra Appl., 93 (1987), pp. 149–162.
[25] S. Van Huffel and J. Vandewalle, *Analysis and solution of the nongeneric total least squares problem*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 327–348.
[26] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem – Computational Aspects and Analysis*, SIAM, Philadelphia, PA, 1991.
[27] S. Van Huffel, J. Vandewalle, and A. Haegemans, *An efficient and reliable algorithm for computing the singular subspace of a matrix, associated with its smallest singular values*, J. Comput. Appl. Math., 19 (1987), pp. 313–330.
[28] S. Van Huffel, and H. Zha, *An efficient total least squares algorithm based on a rank-revealing two-sided orthogonal decomposition*, Numer. Algorithms, 4 (1993), pp. 101–133.
[29] M. Wei, *The analysis for the total least squares problem with more than one solution*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 746–763.
[30] M. Wei, *Algebraic relations between the total least squares and least squares problems with more than one solution*, Numer. Math., 62 (1992), pp. 123–148.