REGULARIZATION BY TRUNCATED TOTAL LEAST SQUARES*

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

Abstract. The Total Least Square (TLS) method is used successfully as a 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 contaminated by errors. This paper focuses on the use of TLS for solving problems with very ill-conditioned coefficient matrices (so-called discrete ill-posed problems), where some regularization is necessary to stabilize the computed solution. In particular, we propose a truncated TLS method in which the small singular values are discarded, discuss the regularizing properties of this method, and present an iterative algorithm based on Lanczos bidiagonalization.

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

1. Introduction.

The Total Least Squares (TLS) method is a technique for solving overdetermined linear systems of equations. It was independently derived in several literatures, 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 [25, 26, 27], and this literature has advanced the algorithmic and theoretical understanding of the method.

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:

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

In contrast to this, the ordinary Least Squares (LS) method requires that $\tilde{A} = A$, and minimizes the 2-norm of the residual vector $b - \tilde{b}$.

Recently, Fierro and Bunch [5] extended the TLS technique to problems where the matrix A is *nearly rank deficient*, i.e., where A has one or more small singular values, and where there is a well-defined gap between the large and small singular values of A. Their idea is to simply ignore all the small singular values of $(A\,,\,b)$ and treat the problem as an exactly rank-deficient one. 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, where small singular values of A are ignored. In both methods, the almost redundant information in $(A\,,\,b)$ and A, respectively, associated with the small singular values, is discarded and the original

* The work of R. D. Fierro and P. C. Hansen was partially sponsored by a NATO Collaborative Research Grant 5-2-05/RG900098. The work of G. H. Golub 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 No. F49620–91–C–0086. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyrights notation hereon.

† Department of Mathematics, California State University, San Marcos, CA 92096 (fierro@thunder.csusm.edu).

‡ Department of Computer Science, Stanford University, Stanford, CA 94305 (golub@sccm.stanford.edu).

§ UNI•C (Danish Computing Center for Research and Education), Building 305, Technical University of Denmark, DK-2800 Lyngby, Denmark (Per.Christian.Hansen@uni-c.dk).

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

ill-conditioned problem is replaced with another near-by 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 also 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 [6]. 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 to problems with well-defined numerical rank has been given by Wei [30, 31].

There are also many ill-conditioned problems arising in practical applications for which A does not have a well-determined numerical rank, and instead its 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] in order to suppress the effects of errors.

Most regularization methods used today assume that the errors in $A\,x \approx b$ are confined to the right-hand side. Although this is true in many applications there are also problems in which both A and b are contaminated by errors. 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.

The purpose of this paper is to investigate the truncated TLS technique and show that it produces a regularized 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 to the dimensions of A.

The basis for our analysis is the singular value decomposition 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, and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ with $\sigma_1 \geq \cdots \geq \sigma_n$. Then the instabilities associated with discrete ill-posed problems can easily be illustrated. Consider the ordinary LS solution, which can be written as

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

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

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

showing that this approach suppresses the components of the solution corresponding to the small singular values of A, see; e.g., [12, §5.1] or [16]. In this paper we prove that the same is true for truncated TLS.

A fundamental concept that needs to be mentioned here is the *discrete Picard condition* [15]. This criterion states that the coefficients $u_i^T b^{\mathrm{exact}}$ associated with the unperturbed right-hand side b^{exact} must, on average, decay faster than A's singular values—otherwise regularization does not lead to a stabilized solution.

Our paper is organized as follows. Section 2 introduces our main idea, the truncated TLS algorithm, and the regularizing properties of this algorithm are analyzed in §3 and §4. In §5 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 §6. Finally, in §7 we present numerical results. We do not address the important issue of scaling of A and b; see instead [27, §3.6.2] for some details.

2. Truncated TLS. We shall first make precise 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 factorizations have also appeared [2, 29]. For clarity, in this section we shall confine ourselves to the SVD-based approach, and return to computational and algorithmic aspects in §5. Given an $m \times n$ matrix A and an m-vector b, the standard TLS procedure is the following [27, §3.6.1]:

Algorithm TLS

1. Compute the SVD of the compound 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 $\bar{\sigma}_1 \geq \ldots \geq \bar{\sigma}_{n+1}$.

2. Determine the *largest* integer p for which

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

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

$$\begin{array}{cc} & \begin{array}{cc} \overset{p}{\longleftrightarrow} & \overset{q}{\longleftrightarrow} \end{array} \\ \bar V = & \left(\begin{array}{cc} \bar V_{11} & \bar V_{12} \\ \bar V_{21} & \bar V_{22} \end{array} \right) \begin{array}{l} \updownarrow \;\; n \\ \updownarrow \;\; 1 \end{array} . \end{array} \tag{8}$$

4. Compute the minimum norm TLS solution $\bar x_p$ as

$$\bar x_p = -\bar V_{12}\, \bar V_{22}^{\dagger} = -\bar V_{12}\, \bar V_{22}^{T}\, \|\bar V_{22}\|_2^{-2} . \tag{9}$$

In (9), $\bar V_{22}^{\dagger}$ denotes the pseudoinverse of $\bar V_{22}$ which is easy to compute because $\bar V_{22}$ is a vector. This algorithm includes the extensions from [27]. In particular, the first condition in step 2 ensures that Algorithm TLS computes the unique minimum norm solution, while the second condition ensures the existence of a solution. Together, the two conditions ensure that $\sigma_p > \bar\sigma_{p+1}$ (in fact, (7) is equivalent to $\sigma_p > \bar\sigma_{p+1}$).

The norms of $\bar x_p$ and the corresponding TLS residual matrix are given by

$$\|\bar x_p\|_2 = \sqrt{\left\|\bar V_{22}\right\|_2^{-2} - 1} \tag{10}$$

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

We see that $\|\bar x_p\|_2$ increases with p while the residual norm decreases with p.

As mentioned in the Introduction, the approach taken in the truncated SVD technique is simply to compute the SVD of the coefficient matrix A, neglect the small singular values, and then solve a modified rank-deficient least squares problem where A is replaced by the rank-k matrix $\sum_{i=1}^{k} u_i\, \sigma_i\, v_i^T$, where k is the number of large singular values that are retained [16]. We take a similar approach in the truncated TLS method by neglecting the small singular values of $(A\,,\,b)$. Thus, we first determine the number k of *large* singular values $\bar\sigma_i$ of $(A\,,\,b)$; for example, we can take k as the number of $\bar\sigma_i$ larger than some user-specified threshold, or k can be determined adaptively; cf. §5.2. Then we form a rank-k approximation to $(A\,,\,b)$ as

$$(\hat A\,,\,\hat b) \equiv \sum_{i=1}^{k} \bar u_i\, \bar\sigma_i\, \bar v_i^T . \tag{12}$$

Finally, we apply Algorithm TLS to the rank-k matrix $(\hat A\,,\,\hat b)$ to compute the minimum norm solution to this problem. We call this solution the *truncated TLS solution*, and we denote it by $\bar x_k$. The complete algorithm for computing the truncated TLS solution thus becomes:

Algorithm T-TLS

> Before Step 2 in Algorithm TLS, choose a truncation parameter k less than rank$(A\,,\,b)$, and set $\bar\sigma_{k+1} = \ldots = \bar\sigma_{n+1} = 0$.

We note that in Algorithm T-TLS the number p can only differ from the truncation parameter k if the TLS problem associated with $(\hat A\,,\,\hat b)$ in (12) is *nongeneric* (cf. [27, §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. In the next section we analyze this situation.

The truncated TLS solution $\bar{x}_k$ computed by means of Algorithm T-TLS can be expressed in terms of the SVD of $(A\,,\,b)$. It follows from [30, Theorem 2.2] that $\bar{x}_k$ satisfies the equation

$$(A^TA - \bar{V}_{12}\,\bar{\Sigma}_2^2\bar{V}_{12}^T)\,\bar{x}_k = A^Tb - \bar{V}_{12}\,\bar{\Sigma}_2^2\bar{V}_{22}^T\ ,$$

where $\bar{\Sigma}_2 = \mathrm{diag}(\bar{\sigma}_{k+1},\ldots,\bar{\sigma}_{n+1})$.

3. Near-Nongeneric TLS Problems. The issue of a small $\|\bar{V}_{22}\|_2$ has not been analyzed in the literature, except for the case when $\|\bar{V}_{22}\|_2 = 0$. Since the size of $\|\bar{V}_{22}\|_2$ plays such an important role in TLS problems—it is a measure of distance to the nearest nongeneric problem [5]—we will analyze it carefully here.

If $\|\bar{V}_{22}\|_2 = 0$ then the truncated TLS solution cannot be computed from (9). However, from [27, Theorem 3.22], $\|\bar{V}_{22}\|_2 = 0$ implies that b is orthogonal to

$$\mathrm{span}\{\bar{u}_{k+1},\ldots,\bar{u}_{n+1}\} = \mathrm{span}\{u_k,\ldots,u_n\}\ .$$

This means that the truncated SVD solutions x_k and x_{k-1} of (3) are identical. Moreover, Step 2 of Algorithm T-TLS ensures that one merely chooses a lower-rank approximation, namely, $\sum_{i=1}^{p}\bar{u}_i\bar{\sigma}_i\bar{v}_i^T$, where $p < k$, and Step 3 effectively derives a T-TLS solution $\bar{x}_p$ from $\mathrm{span}\{\bar{v}_{p+1},\ldots,\bar{v}_k\}$. Then $\bar{x}_p$ is the minimum norm solution to $\hat{A}_p\,x = \hat{b}_p$, where

$$(\hat{A}_p\,,\,\hat{b}_p) = (A\,,\,b)\,\big(I - (\bar{w}_p,\bar{v}_{k+1},\ldots,\bar{v}_{n+1})(\bar{w}_p,\bar{v}_{k+1},\ldots,\bar{v}_{n+1})^T\big)\ ,$$

$\bar{w}_p = (\bar{x}_p^T\,,\,-1)^T(1+\|\bar{x}_p\|_2^2)^{-1/2}$, and $\mathrm{rank}\big((\hat{A}_p\,,\,\hat{b}_p)\big) = \mathrm{rank}(\hat{A}_p) = k-1$. Thus, both truncated SVD and truncated TLS solve equidimensional problems, a point that is elaborated in [27].

However, in general $\|\bar{V}_{22}\|_2 \neq 0$, but it can be arbitrarily small. We will show that the same relationships nearly hold in this situation, i.e., a small $\|\bar{V}_{22}\|_2$ implies that b is nearly orthogonal to $\mathrm{span}\{u_k,\ldots,u_n\} \approx \mathrm{span}\{\bar{u}_{k+1},\ldots,\bar{u}_{n+1}\}$, and the following theorem quantifies this.

Theorem 3.1. *Let $(A\,,\,b)$ have the SVD in (6), and let $\bar{v}_i(1\!:n)$ represent the first n components of $\bar{v}_i$. Then for $i = 1,\ldots,n+1$*

$$\frac{\big\|(A^TA - \bar{\sigma}_i^2I)\,\bar{v}_i(1\!:n)\big\|_2}{\|\bar{v}_i(1\!:n)\|_2} = \frac{\big\|A^Tb\big\|_2\,|\bar{v}_{n+1,i}|}{\sqrt{1-\bar{v}_{n+1,i}^2}} \tag{13}$$

and

$$\big\|(AA^T - \bar{\sigma}_i^2I)\,\bar{u}_i\big\|_2 = \|b\|_2\,\bar{\sigma}_i\,|\bar{v}_{n+1,i}|\ . \tag{14}$$

Proof. To prove (13), the eigenequation $(A\,,\,b)^T(A\,,\,b)\,\bar{v}_i = \bar{\sigma}_i^2\,\bar{v}_i$ implies

$$(A^TA - \bar{\sigma}_i^2I)\,\bar{v}_i(1\!:n) = -A^Tb\;\bar{v}_{n+1,i}\ .$$

Taking norms and using the fact $\|\bar{v}_i(1\!:n)\|_2^2 = 1-\bar{v}_{n+1,i}^2$, the desired result follows. To prove (14) we begin with the eigenequation $(A\,,\,b)(A\,,\,b)^T\,\bar{u}_i = \bar{\sigma}_i^2\,\bar{u}_i$ or, equivalently,

$$(AA^T - \bar{\sigma}_i^2I)\,\bar{u}_i = b\,b^T\bar{u}_i\ .$$

Table 1

The relative differences between the TSVD solutions x_k and the T-TLS solutions $\bar{x}_k$ and the least squares solution x^*, with corresponding residuals for the 4×3 example.

k	$\|\|x^* - x_k\|\|_2/\|\|x^*\|\|_2$	$\|\|b - Ax_k\|\|_2$	$\|\|x^* - \bar{x}_k\|\|_2/\|\|x^*\|\|_2$	$\|\|(A\,,\,b) - (\tilde{A}\,,\,\tilde{b})\|\|_F$
1	$9.99 \cdot 10^{-2}$	1.00	$9.94 \cdot 10^{2}$	1.00
2	$9.95 \cdot 10^{-3}$	1.00	$9.85 \cdot 10^{8}$	$1.00 \cdot 10^{-1}$
3	$1.35 \cdot 10^{-17}$	1.00	$9.95 \cdot 10^{13}$	$1.00 \cdot 10^{-3}$

Now, $b^T \bar{u}_i = \bar{\sigma}_i \, \bar{v}_{n+1,i}$ and the desired result follows by substitution and taking norms. This completes the proof. □

Remark. For $\bar{v}_{n+1,i} = 0$ the results in Theorem 3.1 coincide with the results in [27, Theorem 3.11].

A nice feature in Theorem 3.1 is the equalities in the relationships. From (13) and (14) we see that if $|\bar{v}_{n+1,i}|$ is "small" then $(\bar{\sigma}_i, \bar{u}_i, \bar{v}_i(1\!:n))$ nearly approximates a singular triplet of A. Further, if $||\bar{V}_{22}||_2$ is small then b is nearly orthogonal to $\text{span}\{\bar{u}_{k+1}, \ldots, \bar{u}_n\}$, and $\text{span}\{u_k, \ldots, u_n\} \approx \text{span}\{\bar{u}_{k+1}, \ldots, \bar{u}_{n+1}\}$ provided σ_{k-1} is not too close to σ_k. This means that the truncated SVD solution essentially lies in a lower dimensional subspace, namely, $\text{span}\{v_1, \ldots, v_{k-1}\}$ and that the dimension of the truncated TLS problem should be reduced to stabilize the solution; consequently, both methods produce solutions in equidimensional subspaces. The discrete Picard condition ensures that such a deflation does not alter the truncated SVD solution drastically.

Thus, throughout the rest of this paper we shall assume that $||\bar{V}_{22}||_2$ is not too "small" since we can always decrease k to satisfy this assumption.

One example of a situation that leads to a near-nongeneric TLS problem is a highly incompatible problem $A\,x \approx b$. We can illustrate this by the following small example. Let

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & .1 & 0 \\ 0 & 0 & .001 \\ 0 & 0 & 0 \end{pmatrix}, \qquad b = \begin{pmatrix} 10^{-3} \\ 10^{-5} \\ 10^{-8} \\ 1 \end{pmatrix}, \qquad U^T b = \begin{pmatrix} 10^{-3} \\ 10^{-5} \\ 10^{-8} \end{pmatrix}. \tag{15}$$

The exact LS solution is $x^* = (10^{-3}, 10^{-4}, 10^{-5})^T$ and the system satisfies the discrete Picard condition. The bottom row of $\bar{V}$ in the SVD of $(A\,,\,b)$ is

$$\left(\bar{V}_{21}\ \bar{V}_{22}\right) = (.7028,\, .7069,\, -10^{-6},\, 10^{-11})\ ,$$

and the small elements show that the TLS problem is nearly nongeneric. Table 1 shows the relative differences between the least squares solution x^* and the solutions x_k and $\bar{x}_k$ for $k = 1, 2, 3$. The truncated SVD solutions x_k are all regularized approximations to x^*. The truncated TLS solutions $\bar{x}_k$ give much smaller residuals but are in no sense approximations to x^*.

We remark that such highly incompatible problems are not likely to occur in practice, but the example still illustrates the difficulties associated with a near-nongeneric problem.

4. Regularizing Properties of the Truncated TLS Solution. In this section we take a closer look at the truncated TLS solution $\bar{x}_k$ and show that it is a *regularized* solution.

Fierro and Bunch [5, §3] showed that

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

Hence, if there is a well-defined gap between the large and small singular values of A, and if $\|\bar{x}_k\|_2$ is not too large relative to $\|x_k\|_2$, then the truncated TLS and the truncated SVD solutions are guaranteed to be similar and $\bar{x}_k$ is a regularized solution.

Our present analysis differs from the analysis by Fierro and Bunch because we do not assume a gap in the singular value spectrum. We stress that we still 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; cf. (10).

We start with an important theorem which relates the truncated TLS solution $\bar{x}_k$ to the SVD of A—and not to the SVD of $(A\,,\,b)$ as is common in the literature.

THEOREM 4.1. *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 are distinct. Write the truncated TLS solution* $\bar{x}_k$ *in the form*

$$\bar{x}_k = \sum_{i=1}^{n} f_i\,\frac{u_i^T b}{\sigma_i}\,v_i\ , \tag{16}$$

where f_i *are the filter factors for truncated TLS. Then the filter factors are given by*

$$f_i = \sum_{j=k+1}^{n+1} \frac{\bar{v}_{n+1,j}^2}{\left\|\bar{V}_{22}\right\|_2^2}\left(\frac{\sigma_i^2}{\sigma_i^2-\bar{\sigma}_j^2}\right)\quad ,\quad i=1,\ldots,n\ . \tag{17}$$

If $\bar{\sigma}_j = \sigma_i$ *for some j then the corresponding term does not contribute to* f_i*.*

Proof. The theorem is proved by considering the updating of the SVD of A when b is appended. It is shown in [1, §4.2] that the columns $\bar{v}_j$ of $\bar{V}$ for which $\bar{v}_{n+1,j}$ are nonzero are given by

$$\bar{v}_j = \bar{w}_j/\|\bar{w}_j\|_2$$

where

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

We see from (9) that these are the only columns that contribute to the solution. Moreover, it is proved in [27, Theorem 3.11] that $\bar{v}_{n+1,j} = 0 \Leftrightarrow \bar{\sigma}_j = \sigma_i$. Eq. (17) then follows immediately by inserting the above expressions into (9). □

REMARK. If $k = n$ or $\bar{\sigma}_{k+1} = \cdots = \bar{\sigma}_{n+1}$, then (17) reduces to $f_i = \sigma_i^2/(\sigma_i^2 - \bar{\sigma}_{n+1}^2)$, consistent with [27, Theorem 2.7].

We shall now give a further characterization of the filter factors for truncated TLS and thus show that $\bar{x}_k$ is indeed a regularized solution. Because of step 2 in ALGORITHM TLS, we can assume without loss of generality that $\sigma_k \neq \bar{\sigma}_{k+1}$.

THEOREM 4.2. *Assume that* $\sigma_k \neq \bar{\sigma}_{k+1}$*. Then the first k filter factors* f_i *form a monotonically increasing sequence and satisfy*

$$0 \leq f_i - 1 \leq \frac{\bar{\sigma}_{k+1}^2}{\sigma_i^2 - \bar{\sigma}_{k+1}^2}\ ,\qquad i=1,\ldots,k \tag{18}$$

while the last $n-k$ filter factors satisfy

$$(19) \qquad 0 \le f_i \le \left\|\bar{V}_{22}\right\|_2^{-2} \frac{\sigma_i^2}{\bar{\sigma}_k^2 - \sigma_i^2} \ , \qquad i = k+1, \ldots, n \ .$$

Proof. To prove (18) we have for $i = 1, \ldots, k$

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

It follows from the interlacing inequalities for the singular values of A and $(A\,,\,b)$ that $\sigma_i \ge \bar{\sigma}_{k+1}$ for $i = 1, \ldots, k$. Hence, with the assumption $\sigma_k \neq \bar{\sigma}_{k+1}$, the second term in the above equation for f_i is positive and we have proved the left inequality in (18). The right inequality follows from the facts that $\sum_{j=k+1}^{n+1} \bar{v}_{n+1,j}^2 = \left\|\bar{V}_{22}\right\|_2^2$ and $\bar{\sigma}_j^2/(\sigma_i^2 - \bar{\sigma}_j^2) \le \bar{\sigma}_{k+1}^2/(\sigma_i^2 - \bar{\sigma}_{k+1}^2)$ for $j = k+1, \ldots, n+1$.

The proof for (19) is based on the secular equations associated with downdating the SVD of $(A\,,\,b)$ when b is deleted [1, §5]:

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

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

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

By means of the relation $\bar{\sigma}_j^2/(\bar{\sigma}_j^2 - \sigma_i^2) = 1 + \sigma_i^2/(\bar{\sigma}_j^2 - \sigma_i^2)$, and using the fact that $\bar{v}_{n+1,j}^2$ sum to one, we can rewrite the secular equations as

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

Using this relation and Eq. (17) for the filter factors, we have for $i = k+1, \ldots, n$:

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

Then the interlacing inequalities for singular values and the assumption $\sigma_k \neq \bar{\sigma}_{k+1}$ ensure that f_i is positive. Finally, using the relation $\sum_{j=1}^{k} \bar{v}_{n+1,j}^2 = \left\|\bar{V}_{21}\right\|_2^2 \le 1$, we obtain

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

Thus, we have proved (19). □

COROLLARY 4.3. *The norms of $\bar{x}_k$ and x_k satisfy*

$$\textit{(20)} \qquad \|\bar{x}_k\|_2 \geq \|x_k\|_2 \ , \qquad k = 1, \ldots, n \ .$$

Proof. Equation (20) 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 1 and 0. □

From Theorem 4.2 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 f_i is to 1. 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 0. Hence, Theorem 4.2 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 gap in the singular value spectrum, provided that $\|\bar{V}_{22}\|_2$ is not very small.

We conclude that if the discrete Picard condition is satisfied, then the truncated TLS solution $\bar{x}_k$ is a regularized 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 $\bar{x}_k$. (For more details why such a solution is a regularized solution we refer to the analysis of the truncated SVD solution in [16].) Moreover, we see that the truncation parameter k plays the role of a regularization parameter.

The difference between the truncated TLS solution and the truncated SVD solution is due to the fact that the truncated TLS technique takes into account the errors in the coefficient matrix A. If $k = n$, then the difference $\|x_k - \bar{x}_k\|_2$ is sometimes quite small, in particular when the errors in A and b are small [24]. When $k < n$, the examples in [5, 6], as well as our examples in §7, illustrate that x_k and $\bar{x}_k$ can be very different.

5. 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 more suited for large-scale problems whenever $k \ll n$, which is indeed the case in most discrete ill-posed problems.

A fairly straight-forward 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

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

The partial SVD can be computed by a technique similar to the PSVD algorithm described in [28] 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.

5.1. 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 know 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, 21]. 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 compound matrix $(A\,,\,b)$, after which we could compute approximate truncated TLS solutions for various k less than $k_{\max}$ by means of Eq. (21).

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

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

or

$$\min \|(B_k\,,\,\beta_1 e_1) - (\hat{B}_k\,,\,\hat{e}_k)\|_F \qquad \text{subject to} \qquad \hat{B}_k\, y = \hat{e}_k \ , \tag{22}$$

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 [22] 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 ALGORITHM TLS to the small-size problem in (22). Hence, we compute the SVD of the matrix $(B_k\,,\,\beta_1 e_1)$,

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

and the standard TLS solution $\bar{y}_k$ to (22) 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

$$\tilde{x}_k = -V_k\, \bar{y}_k = -V_k\, \bar{\bar{V}}_{12}^{(k)} \left(\bar{\bar{v}}_{22}^{(k)}\right)^{-1} \ . \tag{23}$$

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 [28].

We remark that it is easy to augment the above algorithm to include the computations of the LSQR algorithm [22]. Approximate TSVD solutions can be computed together with the approximate T-TLS solutions with little extra overhead.

5.2. 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 5.1. *The norms of the solution and the residual matrix in the Lanczos T-TLS algorithm satisfy*

$$\|\tilde{x}_k\|_2 = \sqrt{\left(\bar{\bar{v}}_{22}^{(k)}\right)^{-2} - 1} \tag{24}$$

and

$$\|(A\,,\,b) - (\hat{A}_k\,,\,\hat{b}_k)\|_F^2 = \|(A\,,\,b)\|_F^2 - \|(B_k\,,\,\beta_1 e_1)\|_F^2 + (\bar{\bar{\sigma}}_{k+1}^{(k)})^2\ , \tag{25}$$

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 (25) is a nonincreasing function of k.

Proof. Equations (24) and (25) 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 [22]. 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 Eq. (6:3) of Hestenes and Stiefel [20].

Notice that (25) 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, §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 $\hat{k} = 1, 2, ..., k$, 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 §7.

Another popular method for choosing the regularization parameter is the method of Generalized Cross-Validation due to Wahba [8]. Currently, we do not have any experience with this method.

A third possible stopping criterion can be based on the L-curve criterion studied recently in [17, 19]. 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 [19].

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.

6. Regularization in General Form. Theorems 4.1 and 4.2 show that the T-TLS solution $\bar{x}_k$ is a regularized solution whose main contributions come from the first k right singular vectors v_i. It is common experience 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):

$$A = \breve{U}\,\mathrm{diag}(\alpha_i)\,W^{-1}\ , \qquad L = \breve{V}\,\mathrm{diag}(\beta_i)\,W^{-1}\ , \tag{26}$$

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_1/β_i. See, for example, [14], [13, §4] or [18, §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

$$A_{\mathrm{sf}} = A\,L_A^{\dagger} = \breve{U}\,\mathrm{diag}(\alpha_i/\beta_i)\,\breve{V}^T\ , \qquad b_{\mathrm{sf}} = b - A\,x_0\ , \tag{27}$$

[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 $L_A^\dagger$:

$$x = L_A^\dagger x_{\mathrm{sf}} + x_0 \ . \tag{28}$$

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

```
COMPUTE $y = (L_A^\dagger)^T x$
    1. $x \leftarrow x - T^T N^T x$
    2. $\begin{pmatrix} y \\ z \end{pmatrix} \leftarrow \begin{pmatrix} \multicolumn{2}{c}{L} \\ 0 & I_{n-\ell} \end{pmatrix}^{-T} x$
```

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, §4.3].

7. 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 [18].

Our test problems were generated as follows. The matrix A is 64×32 and comes from discretization of Phillips's test problem (cf. [18, 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 distributed between 1 and 10^4
$\eta_8^{[1]}, \ldots, \eta_{32}^{[1]}$ are geometrically distributed between 10^4 and 10^{-16}.
For the second,
$\eta_1^{[2]}, \ldots, \eta_{32}^{[2]}$ are geometrically distributed between 1 and 10^{-16}
Only $b^{[1]}$ has coefficients $\eta_i^{[1]}$ that increase with i, and from the theory in [5] we therefore expect that TLS is superior to LS for $b^{[1]}$ only. Both systems are scaled such that $\max_{ij} |a_{ij}| = \max_i |b_i^{[p]}| = 1$. 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.

In connection with the tests reported below, we made the interesting observation that when we perturb the matrix A randomly as described above, then the singular vectors of A are perturbed in a very systematic way. 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.

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 §6 using the implementations gen_form and std_form from [18].

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 $x^{\rm opt}$ as the one closest to the exact solution. E.g., for T-TLS,

$$||\bar{x}^{\rm opt} - x^{\rm exact}||_2 \leq ||\bar{x}_k - x^{\rm exact}||_2 \ , \qquad k = 1, \ldots, 12 \ .$$

In this way, we can investigate in 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 8 coefficients $\eta_i^{[1]}$ increase. Figure 1 shows histograms of the relative errors $||x^{\rm opt} - x^{\rm exact}||_2/||x^{\rm exact}||_2$ 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.

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 is that same as in Test 1, the histograms for all five methods are now almost identical. Notice, in particular, the resemblance of T-TLS, Tikhonov and TSVD, and the resemblance of Lanczos T-TLS and LSQR.

Test 3. Our final test problem uses the second right-hand side $b^{[2]}$ (which satisfies the discrete Picard condition for all coefficients) 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 when the discrete Picard condition is violated. 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 discrete Picard condition is satisfied, or when the errors are "small", then there is no advantage in using the TLS approach over the classical methods.

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

For test problem 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 is 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 test problems 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 [17, 19] for numerical examples. Further research in this area is required.

A. Appendix. In this appendix we complete the proof of Theorem 5.1 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 $\left(\bar{\bar{\sigma}}_1^{(k)}\right)^2, \ldots, \left(\bar{\bar{\sigma}}_{k+1}^{(k)}\right)^2$. 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, §9] and that the eigenvector w associated with the smallest eigenvalue s_k^2 has k sign changes [7, p. 105], i.e.,

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

[3] This result can also be established as a consequence of (3.4.8) in Szegö [23], 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.

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*

$$s_k^2 \leq \min_i \tau_i \quad \text{for} \quad k > 0 \ . \tag{29}$$

Proof. We know that $s_k^2 \leq z^T T_k z$ for any vector z of length 1. 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*

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

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

$$w_1 - z_1 \geq 0 \ . \tag{30}$$

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 (30).

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 1st 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

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

so $z_2 < w_2 < 0$. A similar computation for the 2nd row yields

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

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 $s_k^2 = \left(\bar{\bar{\sigma}}_{k+1}^{(k)}\right)^2$ and $s_{k+1}^2 = \left(\bar{\bar{\sigma}}_{k+2}^{(k+1)}\right)^2$ 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 §5.1. Thus, $|\bar{\bar{v}}_{22}^{(k)}| - |\bar{\bar{v}}_{22}^{(k+1)}| = w_1 - z_1 \geq 0$ for all $k > 0$.

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

REFERENCES

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

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

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

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

[5] R. D. Fierro & J. R. Bunch, *Collinearity and total least squares*, Report, Dept. of Mathematics, University of California, San Diego; SIAM J. Matrix Anal. Appl., to appear.

[6] R. D. Fierro & J. R. Bunch, *Perturbation theory for orthogonal projection methods with application to LS and TLS*, CAM Report 92-44, Dept. of Mathematics, UCLA, October 1992; SIAM J. Matrix. Anal. Appl., to appear.

[7] F. R. Gantmacher, *The Theory of Matrices; Vol. II*, Chelsea Publishing Company, New York, 1959.

[8] G. H. Golub, M. T. Heath & G. Wahba, *Generalized Cross-Validation as a method for choosing a good ridge parameter*, Technometrics, 21 (1979), pp. 215–223.

[9] G. H. Golub & W. Kahan, *Calculating the Singular Values and Pseudo-Inverse of a Matrix*, SIAM J. Numer. Anal., **2** (Series B) (1965), 205–224.

[10] G. H. Golub & C. F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal. **17** (1980), 883–893.

[11] G. Golub and J. Welsch, *Calculation of Gauss quadrature rules*, Math. Comp. **23** (1969) 221–230.

[12] C. W. Groetsch, *Inverse Problems in the Mathematical Sciences*, Vieweg, Wiesbaden, 1993.

[13] M. Hanke & P. C. Hansen, *Regularization methods for large-scale problems*, Surv. Math. Ind. **3** (1993), 253–315.

[14] P. C. Hansen, *Regularization, GSVD and truncated GSVD*, BIT **29** (1989), 491–504.

[15] P. C. Hansen, *The discrete Picard condition for discrete ill-posed problems*, BIT **30** (1990), 658–672.

[16] P. C. Hansen, *Truncated SVD solutions to discrete ill-posed problems with ill-determined numerical rank*, SIAM J. Sci. Stat. Comput. **11** (1990), 503–518.

[17] P. C. Hansen, *Analysis of discrete ill-posed problems by means of the L-curve*, SIAM Review **34** (1992), 561–580.

[18] P. C. Hansen, *Regularization Tools: A Matlab Package for Analysis and Solution of Discrete Ill-Posed Problems*; Numerical Algorithms **6** (1994), 1–35.

[19] P. C. Hansen & D. P. O'Leary, *The use of the L-curve in the regularization of discrete ill-posed problems*, Report UMIACS-TR-91-142, Dept. of Computer Science, Univ. of Maryland, October 1991 (23 pages); to appear in SIAM J. Sci. Comput.

[20] M. R. Hestenes & E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Natl. Bureau of Standards **49** (1952), 409–436.

[21] D. P. O'Leary & J. A. Simmons, *A bidiagonalization-regularization procedure for large scale discretizations of ill-posed problems*, SIAM J. Sci. Stat. Comput. **2** (1981), 474–489.

[22] C. C. Paige & M. A. Saunders, *LSQR: an algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software **8** (1982), 43–71.

[23] G. Szegö, *Orthogonal Polynomials*, American Mathematical Society, Providence, Rhode Island, 1939.

[24] G. W. Stewart, *ON the invariance of perturbed null vectors under column scaling*, Numer. Math. **44** (1984), 61–65.

[25] S. Van Huffel & J. Vanderwalle, *Algebraic relationships between classical regression and total least-squares estimation*, Lin. Alg. Appl. **93** (1987), 149–162.

[26] S. Van Huffel & J. Vanderwalle, *Analysis and solution of the nongeneric total least squares problem*, SIAM J. Matrix Anal. Appl. **9** (1988), 327–348.

[27] S. Van Huffel & J. Vanderwalle, *The Total Least Squares Problem – Computational Aspects and Analysis*, SIAM, Philadelphia, 1991.

[28] S. Van Huffel, J. Vanderwalle & A. Haegemans, *An efficient and reliable algorithm for computing the singular subspace of a matrix, associated with its smallest singular values*, J. Comp. Appl. Math. **19** (1987), 313–330.

[29] S. Van Huffel & H. Zha, *An efficient total least squares algorithm based on a rank-revealing two-sided orthogonal decomposition*, Numerical Algorithms **4** (1993), 101–133.

[30] M. Wei, *The analysis for the total least squares problem with more than one solution*, SIAM J. Matrix Anal. Appl. **13** (1992), 746–763.

[31] M. Wei, *Algebraic relations between the total least squares and least squares problems with more than one solution*, Numer. Math. **62** (1992), 123–148.

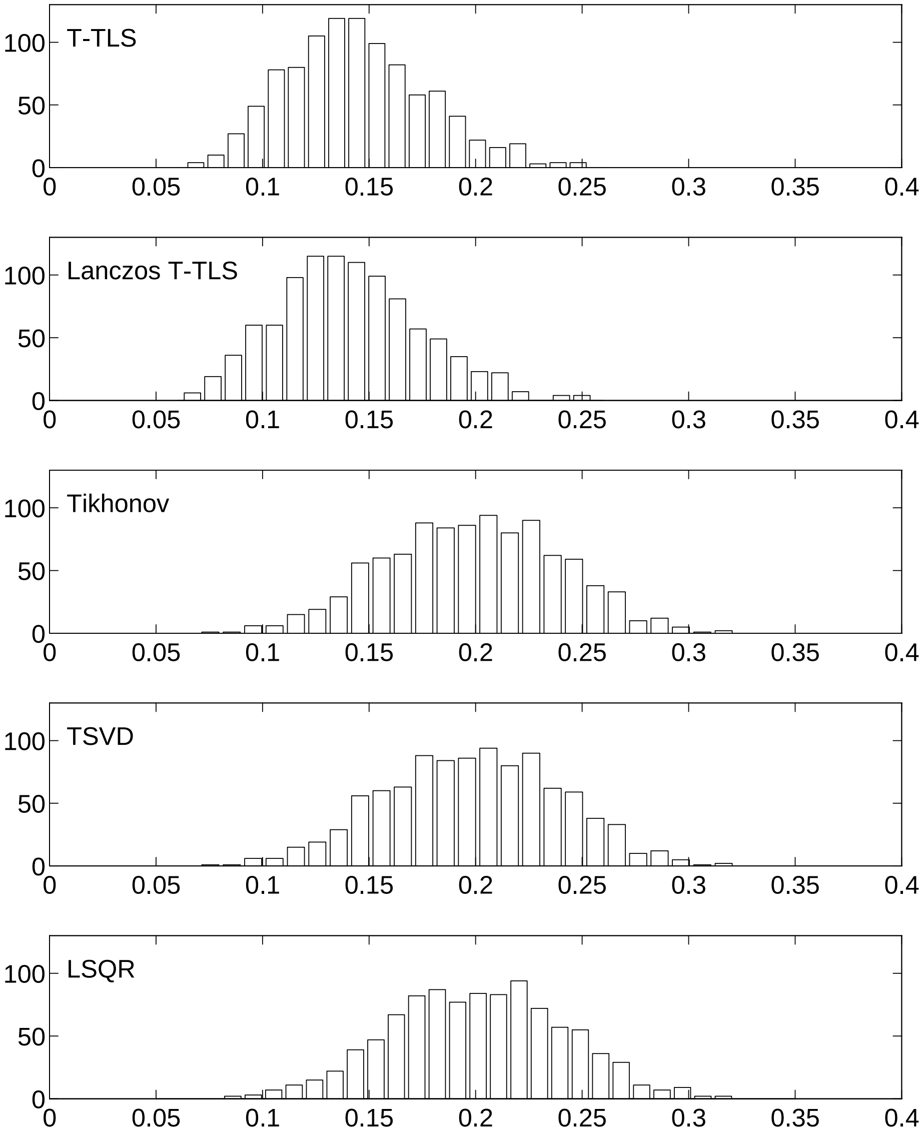

FIG. 1. *Test 1: error level $\epsilon = 5 \cdot 10^{-2}$ and right-hand side $b^{[1]}$. Histograms for the optimal relative errors of 1000 test problems solved by five different regularization methods. Algorithms T-TLS and Lanczos T-TLS are superior to the three classical methods.*

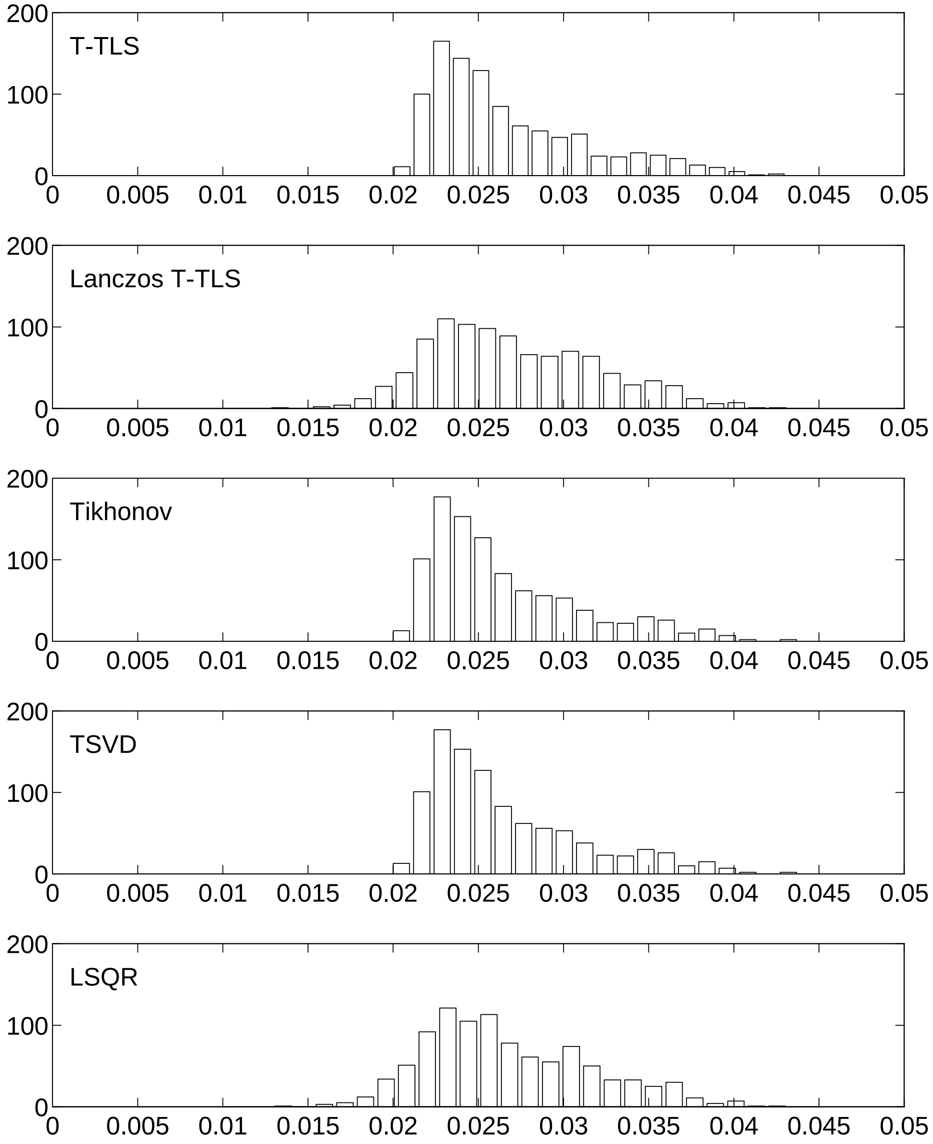

FIG. 2. *Test 2: error level* $\epsilon = 10^{-3}$ *and right-hand side* $b^{[1]}$. *Histograms for the optimal relative errors of 1000 test problems solved by five different regularization methods. All five methods give essentially the same results.*