Springer

Dear Author:

Please find attached the final pdf file of your contribution, which can be viewed using the Acrobat Reader, version 3.0 or higher. We would kindly like to draw your attention to the fact that copyright law is also valid for electronic products. This means especially that:

- You may not alter the pdf file, as changes to the published contribution are prohibited by copyright law.

- You may print the file and distribute it amongst your colleagues in the scientific community for scientific and/or personal use.

- You may make an article published by Springer-Verlag available on your personal home page provided the source of the published article is cited and Springer-Verlag is mentioned as copyright holder. You are requested to create a link to the published article in LINK, Springer's internet service. The link must be accompanied by the following text: The original publication is available on LINK **http://link.springer.de.** Please use the appropriate URL and/or DOI for the article in LINK. Articles disseminated via LINK are indexed, abstracted and referenced by many abstracting and information services, bibliographic networks, subscription agencies, library networks and consortia.

- You are not allowed to make the pdf file accessible to the general public, e.g. your institute/your company is not allowed to place this file on its homepage.

- Please address any queries to the production editor of the journal in question, giving your name, the journal title, volume and first page number.

Yours sincerely,

Springer-Verlag Berlin Heidelberg

Numer. Math. (2002) 91: 93–115
Digital Object Identifier (DOI) 10.1007/s002110100317

Numerische Mathematik

Bounds for the least squares distance using scaled total least squares

Christopher C. Paige[1,⋆], **Zdeněk Strakoš**[2,⋆⋆]

[1] School of Computer Science, McGill University, Montreal, Quebec, Canada H3A 2A7; e-mail: paige@cs.mcgill.ca
[2] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic; e-mail: strakos@cs.cas.cz

Received July 20, 2000 / Revised version received February 28, 2001 /
Published online July 25, 2001 –

Summary. The standard approaches to solving overdetermined linear systems $Bx \approx c$ construct minimal corrections to the data to make the corrected system compatible. In ordinary least squares (LS) the correction is restricted to the right hand side c, while in scaled total least squares (STLS) [14, 12] corrections to both c and B are allowed, and their relative sizes are determined by a real positive parameter γ. As $\gamma \to 0$, the STLS solution approaches the LS solution. Our paper [12] analyzed fundamentals of the STLS problem. This paper presents a theoretical analysis of the relationship between the sizes of the LS and STLS *corrections* (called the LS and STLS distances) in terms of γ. We give new upper and lower bounds on the LS distance in terms of the STLS distance, compare these to existing bounds, and examine the tightness of the new bounds. This work can be applied to the analysis of iterative methods which minimize the residual norm, and the generalized minimum residual method (GMRES) [15] is used here to illustrate our theory.

Mathematics Subject Classification (1991): 15A18, 15A42, 65F10, 65F20, 65F25, 65F50

⋆ Supported by NSERC of Canada Grant OGP0009236.
⋆⋆ Supported by the GA AS CR under grant A1030103. Part of this work was performed during the academic years 1998/1999 and 1999/2000 while visiting Emory University, Atlanta, GA, USA
Correspondence to: Z. Strakoš

1 Introduction

We will use $\mathcal{R}(B)$ to denote the range (column space) of a matrix B. Two useful approaches to solving the overdetermined approximate linear system

$$Bx \approx c, \quad B \text{ an } n \text{ by } k \text{ matrix}, \ c \text{ an } n\text{-vector}, \ \ c \notin \mathcal{R}(B), \tag{1.1}$$

are ordinary least squares (LS, see for example [1], [5, Sect. 5.3]) and scaled total least squares (STLS, presented in [14] where it was called weighted total least squares, and developed further in [12]). STLS is a generalization of total least squares (TLS, see [3,4], and for example [1, Sect. 4.6], [5, Sect. 12.3], [9]).

In LS we seek (we use $\|\cdot\|$ to denote the vector 2-norm)

$$\text{LS distance} \ \equiv \ \min_{r,y} \|r\| \quad \text{subject to} \quad By = c - r. \tag{1.2}$$

In STLS, for a given parameter $\gamma > 0$, z, E and s are sought to minimize the Frobenius (F) norm in

$$\text{STLS distance} \ \equiv \ \min_{s,E,z} \|[s, E]\|_F \quad \text{s. t.} \quad (B + E)z\gamma = c\gamma - s. \tag{1.3}$$

We call the $z = z(\gamma)$ which minimizes this the STLS solution of (1.3). Here the relative sizes of the corrections E and s in B and $c\gamma$ are determined by the real scaling parameter $\gamma > 0$. As $\gamma \to 0$ the STLS solution approaches the LS solution, and when $\gamma = 1$ (1.3) coincides with the TLS formulation. The formulation (1.3) is studied in detail in [12]. In applications γ can have a statistical interpretation, see for example [12, Sect. 1], but here we regard γ simply as a variable.

STLS solutions can be found via the singular value decomposition (SVD). Let $\sigma_{min}(\cdot)$ denote the smallest singular value of a matrix, and let P_k be the orthogonal projector onto the left singular vector subspace of B corresponding to $\sigma_{min}(B)$. This paper will assume

$$\text{the } n \times (k+1) \text{ matrix } [B, c] \text{ has rank } k+1, \text{ and } \ P_k c \neq 0. \tag{1.4}$$

We showed in [12, (3.7)] that this implied

$$0 < \sigma(\gamma) \equiv \sigma_{min}([B, c\gamma]) < \sigma_{min}(B) \quad \text{for all } \gamma > 0. \tag{1.5}$$

In this case the unique solution of the STLS problem (1.3) is obtained from scaling the right singular vector of $[B, c\gamma]$ corresponding to $\sigma_{min}([B, c\gamma])$, and the norm of the STLS correction satisfies, for a given $\gamma > 0$ (see for example [12, (1.9)], or [5, §12.3] when $\gamma = 1$),

$$\text{STLS distance in (1.3)} \ = \ \sigma_{min}([B, c\gamma]). \tag{1.6}$$

This paper is greatly simplified by only dealing with problems where (1.4) holds. The assumption (1.4) is equivalent to that in [12, (1.10)] plus the restriction $c \notin \mathcal{R}(B)$, which eliminates the theoretically trivial case $c \in \mathcal{R}(B)$. It is sufficient to note here that nearly all practical overdetermined problems will already satisfy (1.4), but any overdetermined (and incompatible) problem that does not can be reduced to one that does, see [12, Sect. 8], and the bounds derived here with this assumption will be applicable to the original problem.

It is known that (see for example [12, (6.3)])

$$\lim_{\gamma\to 0} \frac{\text{STLS distance in (1.3)}}{\gamma} = \lim_{\gamma\to 0} \frac{\sigma_{min}([B, c\gamma])}{\gamma} = \|r\|, \text{ the LS distance in (1.2)}, \tag{1.7}$$

but here we examine the relationship between these distances for *any* $\gamma > 0$. This will bound the rate at which these quantities approach each other for small γ, as well as provide bounds on the LS distance in terms of $\sigma_{min}([B, c\gamma])$, and *vice versa*, for all $\gamma > 0$. To facilitate this study we assume (1.4) holds, and introduce the scaled total least squares ratio (STLS ratio) $\tau(\gamma)$ for all $\gamma > 0$, where from (1.7)

$$\tau(\gamma) \equiv \frac{\gamma\|r\|}{\sigma_{min}([B, c\gamma])} \to 1 \text{ as } \gamma \to 0. \tag{1.8}$$

This $\tau(\gamma)$ is the ratio of the LS distance for $Bx \approx c\gamma$ to the STLS distance, and summarizes the relative behaviour of the LS (1.2) and STLS (1.3) distances when (1.4) holds.

Remark 1.1 It will help in reading this paper to realize there are three different items we examine, all as functions of γ:

1. The LS/STLS relationship, via the STLS ratio $\tau(\gamma)$ in (1.8).
2. Bounds on $\|r\|$ in terms of $\sigma_{min}([B, c\gamma])$.
3. Bounds on $\sigma_{min}([B, c\gamma])$ in terms of $\|r\|$.

The second item motivated this study, and is examined at length in Sect. 7, but both the second and third follow from the first. □

Remark 1.2 The case $\gamma = 0$ will either be obvious, for example $\sigma_{min}([B, c\gamma]) = 0$ at $\gamma = 0$, or undefined but with a limit, for example $\tau(\gamma)$ in (1.8). It will in general simplify the presentation to assume $\gamma > 0$, since when $\gamma = 0$ is meaningful, the values will be obvious. □

Van Huffel and Vandewalle [9] derived several useful bounds for TLS versus LS (the $\gamma = 1$ case). Our results extend some of these to the case of general $\gamma > 0$, as well as provide new bounds. This work was initiated by

our research [13] on the finite precision convergence behaviour of GMRES [15]. The results on the second item above are particularly useful for this and also for the analysis of any iterative method which at the step k minimizes $\|Nr_k\|$ for some full column rank matrix N.

The paper is organized as follows. In Sect. 2 we review some mathematical tools that we use, and state the secular equation which $\sigma_{min}([B, c\gamma])$ must satisfy. In Sect. 3 we introduce the important ratio $\delta(\gamma) \equiv \sigma_{min}([B, c\gamma])/\sigma_{min}(B)$, and use the secular equation to prove that when (1.4) holds, $\delta(\gamma)$ is bounded away from unity for all γ. We then study the STLS ratio $\tau(\gamma)$ and the relationship between the bounds on $\tau(\gamma)$, $\|r\|$ and $\sigma_{min}([B, c\gamma])$. In Sect. 4 we use the secular equation to derive such bounds, in particular, bounds on the least squares residual norm $\|r\|$ (LS distance) in terms of the scaled total least squares distance $\sigma_{min}([B, c\gamma])$). We show how good these bounds are, and how varying γ gives important insights into the asymptotic relationship between the LS and scaled TLS distances. In Sect. 5 we compare our bounds to previous results. In Sect. 6 we extend to the case $\gamma \neq 1$ a result from [9], in order to obtain an expression for $\tau(\gamma)$ in terms of the singular values of B and of $[B, c\gamma]$, as well as some related results. In Sect. 7 we briefly discuss the generalized minimum residual method (GMRES) [15], since this is what first motivated the bounds in this paper, and then present numerical results using GMRES to illustrate the theory.

2 Mathematical preliminaries

In this paper we will regularly use the following notation for c and B in (1.1), and for r and y solving the LS problem (1.2). Let $n > k$ in (1.1). Let the $n \times k$ matrix B have rank k and singular values σ_i with singular value decomposition (SVD)

$$B = U_B \Sigma V^H, \quad \Sigma \equiv \text{diag}(\sigma_1, \ldots, \sigma_k), \quad \sigma_1 \geq \ldots \geq \sigma_k > 0. \tag{2.1}$$

Here U_B is $n \times k$ matrix, $U_B^H U_B = I_k$, Σ is $k \times k$, and $k \times k$ V is unitary. Choose a unitary matrix $U = [U_B|\hat{U}_B] = [u_1, \ldots, u_k|u_{k+1}, \ldots, u_n]$ such that $(I - U_B U_B^H)c = u_{k+1}\rho$, $\rho \geq 0$. Then

$$U^H [B, c] \begin{bmatrix} V & 0 \\ 0 & 1 \end{bmatrix} = \left[\begin{array}{cc} \Sigma & a \\ \hline 0 & \rho \\ 0 & 0 \end{array}\right], \tag{2.2}$$

$$a \equiv (\alpha_1, \ldots, \alpha_k)^{\mathrm{T}} \equiv [u_1, \ldots, u_k]^H c = U_B^H c.$$

The elements of a are the components of the vector of observations c in the directions of the left singular vectors of the data matrix B. With (1.2) we

see that

$$U^H r = U^H(c - By) = \begin{bmatrix} a - \Sigma V^H y \\ \rho \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \rho \\ 0 \end{bmatrix}$$

gives the minimum for $\|r\|$. Then for the LS solution and residual

$$y = V\Sigma^{-1}a, \qquad \|y\|^2 = \sum_{i=1}^{k} \frac{|\alpha_i|^2}{\sigma_i^2}, \tag{2.3}$$

$$\|r\| = \rho. \tag{2.4}$$

We will use some classical results. When C in G below is square and nonsingular, the Schur complement (G/C) of C in G is defined below (see for example [8, Sect. 0.8.5]), and the other results follow.

$$(G/C) \equiv F - EC^{-1}D,\ G = \begin{bmatrix} C & D \\ E & F \end{bmatrix} = \begin{bmatrix} I & 0 \\ EC^{-1} & I \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & (G/C) \end{bmatrix} \begin{bmatrix} I & C^{-1}D \\ 0 & I \end{bmatrix},$$

$$\det(G) = \det(C) \cdot \det(G/C). \tag{2.5}$$

For the analysis of the STLS problem (1.3), we will be interested in the singular values σ of $[B, c\gamma]$, see (1.6), and so the eigenvalues σ^2 of $[B, c\gamma]^H[B, c\gamma]$. When (1.4) holds, the smallest singular value of $[B, c\gamma]$ is the STLS distance in (1.3). We now state a useful form of the secular equation for this STLS distance.

Lemma 2.1 *For any $n \times k$ matrix B and n-vector c let $\sigma(\gamma) \equiv \sigma_{min}([B, c\gamma])$. Assume (1.4) holds. Then using the notation in (2.1)–(2.4), $0 < \sigma(\gamma) < \sigma_k \equiv \sigma_{min}(B)$ holds for all $\gamma > 0$, and the STLS distance in (1.3) is $\sigma(\gamma) \equiv \sigma_{min}([B, c\gamma])$, which is the smallest positive solution of*

$$0 = \psi_k(\sigma(\gamma), \gamma) \equiv \det([B, c\gamma]^H[B, c\gamma] - \sigma(\gamma)^2 I) / \det(B^H B - \sigma(\gamma)^2 I)$$

$$= \gamma^2\|r\|^2 - \sigma(\gamma)^2 - \gamma^2\sigma(\gamma)^2 \sum_{i=1}^{k} \frac{|\alpha_i|^2}{\sigma_i^2 - \sigma(\gamma)^2}. \tag{2.6}$$

□

This was derived in [12, Sect. 4]. With $\gamma = 1$, (2.6) was derived in [4], see also [9, Thm. 2.7, & (6.36)]. These latter derivations assumed the weaker condition $\sigma_{min}([B, c]) < \sigma_{min}(B)$, and so do not generalize to STLS for all $\gamma > 0$, see [12].

3 Two useful ratios, $\delta(\gamma)$ and $\tau(\gamma)$

We wish to focus on the relationship between the STLS distance and the LS distance, and its dependence on the scaling parameter γ. The ratio $\delta(\gamma)$ defined next is crucial for the bounds we develop. For any $n \times k$ matrix B of rank k and n–vector c, define for all $\gamma \geq 0$

$$\text{(3.1)}\quad \sigma(\gamma) \equiv \sigma_{min}([B, c\gamma]), \qquad \delta(\gamma) \equiv \sigma_{min}([B, c\gamma])/\sigma_{min}(B).$$

Clearly $\sigma(0) = \delta(0) = 0$. The following lemma on $\delta(\gamma)$ was proven in [12, Corollary 4.1].

Lemma 3.1 *If (1.4) holds and $\gamma > 0$, then $0 < \delta(\gamma) < 1$, and $\delta(\gamma)$ increases as γ increases, and decreases as γ decreases, strictly monotonically.* □

Since we are assuming (1.4) holds, it follows that (1.5) holds, and throughout this paper we have

$$\text{(3.2)}\qquad 0 < \delta(\gamma) \equiv \sigma_{min}([B, c\gamma])/\sigma_{min}(B) < 1 \quad \text{for all } \gamma > 0.$$

We will give several results in terms of the ratio $\delta(\gamma)$. In particular, we will get very tight bounds when $\delta(\gamma)$ is small ($\delta(\gamma) \ll 1$). We see from Lemma 2.1 that when (1.4) holds,

$$\text{(3.3)}\ \sigma_{min}([B, c\gamma]) \leq \gamma\|r\|, \quad \delta(\gamma) \leq \gamma\|r\|/\sigma_{min}(B) \quad \text{for all} \quad \gamma \geq 0,$$

and we can make $\delta(\gamma)$ arbitrarily small by decreasing γ.

Some of our bounds will contain the factor $(1 - \delta(\gamma)^2)^{-1}$, and would be useless if $\delta(\gamma) = 1$ and of limited value when $\delta(\gamma) \approx 1$. We now show that when (1.4) holds, $\delta(\gamma)$ is *bounded away from unity* for all γ, giving an upper bound on $(1 - \delta(\gamma)^2)^{-1}$. It is important that these bounds exist, but remember they are worst case bounds, and give no indication of the sizes of $\delta(\gamma)$ or $(1 - \delta(\gamma)^2)^{-1}$ for the values of γ we will usually be interested in.

Theorem 3.1 *With the notation and assumptions of Lemma 2.1, let $n \times k$ B have singular values $\sigma_1 \geq \ldots \geq \sigma_j > \sigma_{j+1} = \ldots = \sigma_k > 0$. Then since (1.4) holds, and denoting $\beta_k \equiv \|P_k c\|$,*

$$\text{(3.4)}\qquad \beta_k^2 \equiv \|P_k c\|^2 = \sum_{i=j+1}^{k} |\alpha_i|^2 > 0,$$

$$\text{(3.5)}\qquad \delta(\gamma)^2 \equiv \frac{\sigma_{min}^2([B, c\gamma])}{\sigma_k^2} \leq \frac{\|r\|^2}{\beta_k^2 + \|r\|^2} < 1 \ \textit{for all } \gamma \geq 0,$$

$$\text{(3.6)}\qquad (1 - \delta(\gamma)^2)^{-1} \leq 1 + \|r\|^2/\beta_k^2 \ \textit{for all } \gamma \geq 0.$$

Proof. Since (1.4) holds, $\beta_k^2 > 0$ and the minimum positive solution $\sigma(\gamma)$ of (2.6) is $\sigma(\gamma) \equiv \sigma_{min}([B, c\gamma]) < \sigma_k$ for all $\gamma \geq 0$. So

$$\gamma^2\|r\|^2 = \sigma(\gamma)^2 + \frac{\gamma^2\sigma(\gamma)^2\beta_k^2}{\sigma_k^2 - \sigma(\gamma)^2} + \sum_{i=1}^{j} \frac{\gamma^2\sigma(\gamma)^2|\alpha_i|^2}{\sigma_i^2 - \sigma(\gamma)^2} \geq \sigma(\gamma)^2 + \frac{\gamma^2\sigma(\gamma)^2\beta_k^2}{\sigma_k^2 - \sigma(\gamma)^2}.$$

Multiplying by $(\sigma_k^2 - \sigma(\gamma)^2) > 0$ and rearranging gives

$$\sigma(\gamma)^2(\gamma^2\beta_k^2 + \gamma^2\|r\|^2 + \sigma_k^2 - \sigma(\gamma)^2) \leq \gamma^2\|r\|^2\sigma_k^2,$$

$$\delta(\gamma)^2 = \frac{\sigma(\gamma)^2}{\sigma_k^2} \leq \frac{\gamma^2\|r\|^2}{\gamma^2\beta_k^2 + \gamma^2\|r\|^2 + \sigma_k^2 - \sigma(\gamma)^2} \leq \frac{\|r\|^2}{\beta_k^2 + \|r\|^2} < 1, \tag{3.7}$$

proving (3.5), from which (3.6) follows. □

This shows that when (1.4) holds, $\delta(\gamma)$ is bounded away from unity, so $\sigma_{min}([B, c\gamma])$ is bounded away from $\sigma_{min}(B)$, for all γ.

The inequality (3.5) has a useful explanatory purpose. We cannot have $\delta(\gamma) \approx 1$ unless $P_k c$, the projection of c onto the left singular vector subspace of B corresponding to $\sigma_{min}(B)$, is very small compared to r. When $P_k c$ is small compared to r then replacing B by

$$B - \sum_{i=j+1}^{k} u_i \sigma_{min}(B) v_i^H$$

in (1.2) would not increase the LS distance significantly. This confirms that the criterion (1.4) in [12] is exactly what is needeed.

We now return to our second ratio, the STLS ratio $\tau(\gamma)$ in (1.8).

Lemma 3.2 *With the notation and assumptions of Lemma 2.1, let* $\tau(\gamma) \equiv \gamma\|r\|/\sigma_{min}([c\gamma, B])$. *Since (1.4) holds,*

$$1 < \tau(\gamma)^2 = 1 + \gamma^2 \sum_{i=1}^{k} \frac{|\alpha_i|^2}{\sigma_i^2 - \sigma_{min}^2([c\gamma, B])} \to 1 \text{ as } \gamma \to 0.$$

Proof. The equality follows from Lemma 2.1, while (3.5) shows each denominator is bounded away from zero, so the inequality and limit follow. □

This STLS–LS relationship $\tau(\gamma) \to 1$ (see (1.7)) has been presented and proven in [14] for an earlier form of STLS, and in [12, (1.6), (6.3)] for the form (1.3) (neither giving any quantitative results).

We pointed out in Remark 1.1 that we are interested in three items, $\tau(\gamma)$, $\|r\|$ and $\sigma(\gamma)$, and their bounds. Since some results for $\|r\|$ and $\sigma(\gamma)$ can be found from those for $\tau(\gamma)$, it will simplify the paper if we derive some

general relationships here, and apply them later. We will use λ for lower bounds and μ for upper, as in

$$1 \le \lambda_\tau \le \tau(\gamma) \le \mu_\tau, \quad \lambda_r \le \|r\| \le \mu_r, \quad \lambda_\sigma \le \sigma(\gamma) \le \mu_\sigma, \tag{3.8}$$

where for brevity the dependence of the bounds on γ is implied. Bounds derived for $\tau(\gamma)$ give bounds on $\|r\|$ or on $\sigma_{min}([B, c\gamma])$ in the obvious way, as we now document.

Lemma 3.3 *With* $\sigma \equiv \sigma(\gamma) \equiv \sigma_{min}([B, c\gamma]) > 0, \tau(\gamma) \equiv \gamma\|r\|/\sigma, \lambda_\tau \ge 1$ *and* $\gamma > 0$,

$$\{\lambda_\tau \le \tau(\gamma) \le \mu_\tau\} \Leftrightarrow \left\{\lambda_r \equiv \frac{\lambda_\tau \sigma}{\gamma} \le \|r\| \le \mu_r \equiv \frac{\mu_\tau \sigma}{\gamma}\right\}$$
$$\Leftrightarrow \left\{\lambda_\sigma \equiv \frac{\gamma\|r\|}{\mu_\tau} \le \sigma \le \mu_\sigma \equiv \frac{\gamma\|r\|}{\lambda_\tau}\right\}. \quad \square$$

We can easily derive a *single* upper bound on the relative gaps between each pair of upper and lower bounds in Lemma 3.3.

Lemma 3.4 *In Lemma 3.3 the relative gaps between upper and lower bounds satisfy*

$$(\mu_\tau - \lambda_\tau)/\tau \le (\mu_\tau - \lambda_\tau)/\lambda_\tau,$$
$$(\mu_r - \lambda_r)/\|r\| = (\mu_\tau - \lambda_\tau)/\tau \le (\mu_\tau - \lambda_\tau)/\lambda_\tau,$$
$$(\mu_\sigma - \lambda_\sigma)/\sigma = (\mu_\tau - \lambda_\tau)\tau/(\mu_\tau\lambda_\tau) \le (\mu_\tau - \lambda_\tau)/\lambda_\tau. \quad \square$$

The advantage here is that if we can find bounds for $\tau(\gamma)$ such that $(\mu_\tau - \lambda_\tau)/\lambda_\tau$ is sufficiently small, than we can conclude that *all* the bounds in (3.8) are good. This is useful for simplifying the paper.

Lemma 3.2 gave an explicit expression for $0 < \tau(\gamma)^2 - 1 \to 0$ as $\gamma \to 0$. It seems natural to start with this result and to obtain bounds

$$0 \le \lambda \le \tau(\gamma)^2 - 1 \le \mu\,. \tag{3.9}$$

Then the bounds for $\tau(\gamma)$ can simply be given as

$$\lambda_\tau \equiv (\lambda + 1)^{\frac{1}{2}} \le \tau(\gamma) \le \mu_\tau \equiv (\mu + 1)^{\frac{1}{2}} \tag{3.10}$$

and the bounds on $\|r\|$ and σ determined via Lemma 3.3. The relative gap of the τ bounds (showing via Lemma 3.4 how good all these bounds are) is examined in the following lemma.

Lemma 3.5 *For* $\gamma > 0$ *and* $\tau \equiv \tau(\gamma)$, *if we have bounds (3.9) and (3.10), then*

$$\frac{\lambda + \tau + 1}{\tau + 1} \le \lambda_\tau \ \le \ \mu_\tau \ \le \frac{\mu + \tau + 1}{\tau + 1}, \tag{3.11}$$

$$\frac{\mu_\tau - \lambda_\tau}{\tau} \le \frac{\mu_\tau - \lambda_\tau}{\lambda_\tau} \le \frac{\mu - \lambda}{\lambda + \tau + 1} < \frac{\mu - \lambda}{2 + \lambda} < \frac{\mu - \lambda}{2}. \tag{3.12}$$

Proof. With (3.9),

$$\begin{aligned}(\mu+\tau+1)^2-(\tau+1)^2(\mu+1) &= \mu^2+2\mu(\tau+1)-\mu(\tau+1)^2\\ &= \mu[\mu-(\tau^2-1)]\geq 0,\\ (\lambda+\tau+1)^2-(\tau+1)^2(\lambda+1) &= \lambda^2+2\lambda(\tau+1)-\lambda(\tau+1)^2\\ &= \lambda[\lambda-(\tau^2-1)]\leq 0,\end{aligned}$$

proving (3.11). Using (3.11),

$$\frac{\mu_\tau-\lambda_\tau}{\lambda_\tau}\leq\frac{\mu-\lambda}{(\tau+1)\lambda_\tau}\leq\frac{\mu-\lambda}{\lambda+\tau+1}.$$

The rest of the proof is straightforward. □

The actual bounds λ, μ, and, consequently, λ_τ, μ_τ, λ_r, μ_r, λ_σ and μ_σ are developed in the following section.

4 The basic bounds

We have found the following results relating the LS distance $\|r\|$ with the STLS distance $\sigma_{min}([B,c\gamma])$ to be very useful.

Theorem 4.1 *Given a scalar $\gamma>0$, and an n by $k+1$ matrix $[B,c]$, use $\sigma(\cdot)$ to denote singular values and $\|\cdot\|$ to denote 2-norms. If r and y solve $\min_{r,y}\|r\|$ subject to $By=c-r$, and (1.4) holds, then*

$$0<\theta(\gamma)\equiv\frac{\sigma_{min}([B,c\gamma])}{\sigma_{max}(B)}\leq\delta(\gamma)\equiv\frac{\sigma_{min}([B,c\gamma])}{\sigma_{min}(B)}<1, \tag{4.1}$$

and we have bounds on $\tau(\gamma)^2-1$ where $\tau(\gamma)\equiv\gamma\|r\|/\sigma_{min}([B,c\gamma])$:

$$\gamma^2\|y\|^2<\frac{\gamma^2\|y\|^2}{1-\theta(\gamma)^2}\leq\tau(\gamma)^2-1=\frac{\gamma^2\|r\|^2}{\sigma^2_{min}([B,c\gamma])}-1\leq\frac{\gamma^2\|y\|^2}{1-\delta(\gamma)^2}. \tag{4.2}$$

We also have individual bounds on $\tau(\gamma)$, $\|r\|$ and $\sigma_{min}([B,c\gamma])$:

$$\begin{aligned}\lambda_\tau\equiv\{1+\gamma^2\|y\|^2\}^{\frac12} &< \{1+\frac{\gamma^2\|y\|^2}{1-\theta(\gamma)^2}\}^{\frac12}\\ \leq\ \tau(\gamma)\equiv\frac{\gamma\|r\|}{\sigma_{min}([B,c\gamma])} &\leq \mu_\tau\equiv\{1+\frac{\gamma^2\|y\|^2}{1-\delta(\gamma)^2}\}^{\frac12},\end{aligned} \tag{4.3}$$

$$\begin{aligned}\lambda_r\equiv\sigma_{min}([B,c\gamma])\{\gamma^{-2}+\|y\|^2\}^{\frac12} &< \sigma_{min}([B,c\gamma])\,\{\gamma^{-2}+\frac{\|y\|^2}{1-\theta(\gamma)^2}\}^{\frac12}\\ \leq\ \|r\|\leq\mu_r\equiv\sigma_{min}([B,c\gamma])\,\{\gamma^{-2}+\frac{\|y\|^2}{1-\delta(\gamma)^2}\}^{\frac12},\end{aligned} \tag{4.4}$$

$$\lambda_\sigma \equiv \|r\|/\{\gamma^{-2} + \frac{\|y\|^2}{1-\delta(\gamma)^2}\}^{\frac{1}{2}} \ \leq \ \sigma_{min}([B, c\gamma])$$

$$(4.5) \ \leq \|r\|/\{\gamma^{-2} + \frac{\|y\|^2}{1-\theta(\gamma)^2}\}^{\frac{1}{2}} \leq \mu_\sigma \equiv \|r\|/\{\gamma^{-2} + \|y\|^2\}^{\frac{1}{2}}.$$

Proof. Since (1.4) holds, (4.1) follows immediately from (3.2). With $\sigma(\gamma) \equiv \sigma_{min}([B, c\gamma])$ and $\tau(\gamma) \equiv \gamma\|r\|/\sigma(\gamma)$, we see from Lemma 3.2,

$$\sum_{i=1}^{k} \frac{\gamma^2|\alpha_i|^2}{\sigma_i^2} < \sum_{i=1}^{k} \frac{\gamma^2|\alpha_i|^2}{\sigma_i^2(1-\sigma(\gamma)^2/\sigma_1^2)} \leq \frac{\gamma^2\|r\|^2}{\sigma(\gamma)^2} - 1 = \tau(\gamma)^2 - 1 =$$

$$\sum_{i=1}^{k} \frac{\gamma^2|\alpha_i|^2}{\sigma_i^2 - \sigma(\gamma)^2} = \sum_{i=1}^{k} \frac{\gamma^2|\alpha_i|^2}{\sigma_i^2(1-\sigma(\gamma)^2/\sigma_i^2)} \leq \sum_{i=1}^{k} \frac{\gamma^2|\alpha_i|^2}{\sigma_i^2(1-\sigma(\gamma)^2/\sigma_k^2)}$$

(4.6)

which are apparently new bounds on $\tau(\gamma)^2 - 1$. Using (2.3) and (4.1) shows (4.6) is (4.2). We obtain (4.3) by adding 1 to each term in (4.2) and then taking the square root of each term. Both (4.4) and (4.5) follow directly from (4.3), see Lemma 3.3. □

By subtracting $\gamma^2\|y\|^2$ from each term in (4.2), and then dividing by $\gamma^2\|y\|^2$ we obtain bounds of a particularly simple form.

Corollary 4.1 *With the conditions and assumptions of Theorem 4.1*

$$(4.7) \ 0 < \theta(\gamma)^2 < \frac{\theta(\gamma)^2}{1-\theta(\gamma)^2} \leq \ \frac{\tau(\gamma)^2 - (1+\gamma^2\|y\|^2)}{\gamma^2\|y\|^2} \leq \frac{\delta(\gamma)^2}{1-\delta(\gamma)^2}.$$

□

We will now examine the *tightness* of the bounds (4.3)–(4.5), to indicate just how good they can be. In fact we will show that *all* the relative gaps go to zero (as functions of the scaling parameter γ) at least as fast as $O(\gamma^4)$.

Corollary 4.2 *Under the same conditions as in Theorem 4.1, with* $\sigma \equiv \sigma(\gamma) \equiv \sigma_{min}([B, c\gamma])$, $\tau \equiv \tau(\gamma)$, *the notation in (4.3)–(4.5), and*

$$\eta_\tau \equiv (\tau - \lambda_\tau)/\tau, \quad \eta_r \equiv (\|r\| - \lambda_r)/\|r\|, \quad \eta_\sigma \equiv (\sigma - \lambda_\sigma)/\sigma,$$

$$(4.8) \ \zeta_\tau \equiv (\mu_\tau - \lambda_\tau)/\tau, \quad \zeta_r \equiv (\mu_r - \lambda_r)/\|r\|, \quad \zeta_\sigma \equiv (\mu_\sigma - \lambda_\sigma)/\sigma,$$

we have the following bounds

$$0 < \eta_\tau \leq \zeta_\tau, \qquad 0 < \eta_r \leq \zeta_r, \qquad 0 < \eta_\sigma \leq \zeta_\sigma,$$

$$(4.9) \ 0 < \zeta_\tau, \zeta_r, \zeta_\sigma < \frac{\gamma^2\|y\|^2}{2+\gamma^2\|y\|^2} \cdot \frac{\delta(\gamma)^2}{1-\delta(\gamma)^2} \to 0 \ \textit{as} \ \gamma \to 0,$$

where the upper bound goes to zero at least as fast as $O(\gamma^4)$.

Proof. Since $\mu_r, \lambda_r, \mu_\sigma$ and λ_σ were derived from μ_τ and λ_τ as in Lemma 3.3, we see from Lemma 3.4 that ζ_τ, ζ_r and ζ_σ are all bounded above by $(\mu_\tau - \lambda_\tau)/\lambda_\tau$. But λ_τ and μ_τ were obtained directly from λ and μ. Based on (4.2) the bounds λ and μ from (3.9) can be set to $\lambda \equiv \gamma^2\|y\|^2$ and $\mu \equiv \gamma^2\|y\|^2/(1-\delta(\gamma)^2)$. It follows from (3.12) that

$$\frac{\mu_\tau - \lambda_\tau}{\lambda_\tau} \leq \frac{\mu - \lambda}{2+\lambda} = \frac{\gamma^2\|y\|^2}{2+\gamma^2\|y\|^2} \cdot \frac{\delta(\gamma)^2}{1-\delta(\gamma)^2},$$

proving the inequalities. Applying (3.3) shows the γ^4 behaviour. □

Thus when $\delta(\gamma) \ll 1$, or γ is small, the upper and lower bounds in (4.3)–(4.5) are not only very good, but very good in a *relative* sense, which is important for small $\|r\|$ or $\sigma_{min}([B, c\gamma])$. We see Corollary 4.2 makes precise a nice theoretical observation with practical consequences — small γ ensures very tight bounds (4.4) on $\|r\|$. In particular, for small γ we see

$$\|r\| \;\approx\; \lambda_r \equiv \sigma_{min}([B, c\gamma])\,\{\gamma^{-2} + \|y\|^2\}^{\frac{1}{2}}, \tag{4.10}$$

and the relative error is bounded above by $O(\gamma^4)$. Using (4.4) and (3.3) we get another formulation of this result

$$0 < \frac{\|r\|^2 - \sigma_{min}^2([B, c\gamma])(\gamma^{-2} + \|y\|^2)}{\|r\|^2} \leq \frac{\sigma_{min}^2([B, c\gamma])\|y\|^2\delta(\gamma)^2}{\|r\|^2(1-\delta(\gamma)^2)} \tag{4.11}$$

$$\leq \frac{\gamma^2\|y\|^2\delta(\gamma)^2}{(1-\delta(\gamma)^2)} \leq \frac{\gamma^4\|r\|^2\|y\|^2}{\sigma_{min}^2(B)(1-\delta(\gamma)^2)}. \tag{4.12}$$

A crucial aspect of Theorem 4.1 is that it gives both an upper and a lower bound on the minimum residual norm $\|r\|$, or on $\sigma_{min}([B, c\gamma])$, which is the STLS distance in (1.3). The weaker lower bound in (4.4), or upper bound in (4.5), is sufficient for many uses, and is relatively easy to derive, but the upper bound in (4.4), or lower bound in (4.5), is what makes the theorem so strong.

Remark 4.1 When $\delta(\gamma) < 1$, [9, Thm. 2.7] showed (for $\gamma = 1$) the closed form TLS solution $z\gamma = z(\gamma)\gamma$ of (1.3) is

$$z(\gamma)\gamma = \{B^H B - \sigma_{min}^2([B, c\gamma])I\}^{-1} B^H c\gamma,$$

and with $r_{STLS} \equiv c\gamma - Bz(\gamma)\gamma$, [9, (6.19)] showed (for $\gamma = 1$)

$$\|r_{STLS}\| = \sigma_{min}([B, c\gamma])(1 + \|z(\gamma)\gamma\|^2)^{\frac{1}{2}}. \tag{4.13}$$

Relation (4.10) can be seen to give an analogue of this for the LS solution: since $r\gamma = c\gamma - By\gamma$ in (1.2), (4.11) and (4.12) show a strong relationship between $\gamma\|r\|$ and $\sigma_{min}([B, c\gamma])$ for small $\delta(\gamma)$, γ, $\|y\|$ or $\|r\|$:

$$\gamma\|r\| \;\approx\; \sigma_{min}([B, c\gamma])\,\{1 + \gamma^2\|y\|^2\}^{\frac{1}{2}}. \tag{4.14}$$

□

Remark 4.2 It is often useful but difficult to find (a lower bound to) the smallest singular value of a linear operator or a large sparse matrix. These bounds may help. For example suppose we want a lower bound on $\sigma_{min}(B)$. If we can solve some LS problem (1.2) then we know from (4.5) that

$$\|r\|/\{\gamma^{-2} + \frac{\|y\|^2}{1-\delta(\gamma)^2}\}^{\frac{1}{2}} \;\leq\; \sigma_{min}([B, c\gamma]) \;\leq\; \sigma_{min}(B).$$

However we have not looked further into using such bounds. □

Remark 4.3 The assumption $P_k c \neq 0$ in (1.4) is not necessary for proving the bounds (4.2)–(4.5). From the proof of Theorem 4.1 it is clear that these bounds only require $\delta(\gamma) < 1$. However $\delta(\gamma) < 1$ does not guarantee $P_k c \neq 0$. When $P_k c = 0$, $\|r\|$ contains no information whatsoever about $\sigma_{min}(B)$, while the bounds do, see (4.6). By assuming $P_k c \neq 0$ we avoid this inconsistency. Moreover, we will consider various values of the parameter γ, and so we prefer the theorem's assumption to be independent of γ. □

We end this section by a comment on possible consequences of Theorem 4.1 for understanding methods for large TLS problems. It can be shown that the STLS distance $\sigma_{k+1}([B, c\gamma])$ can be analysed via Rayleigh quotients for $[B, c\gamma]^H[B, c\gamma]$:

$$\sigma^2_{k+1}([B, c\gamma]) = \|[B, c\gamma] \begin{pmatrix} -z\gamma \\ 1 \end{pmatrix} \|^2 / \| \begin{pmatrix} -z\gamma \\ 1 \end{pmatrix} \|^2 = \frac{\gamma^2\|c - Bz\|^2}{1+\gamma^2\|z\|^2}$$

where z solves (1.3), see [2]. For small $\sigma_{min}([B, c\gamma])$, $\delta(\gamma)$, γ or $\|y\|$, (4.5) with (4.9) show

$$\sigma^2_{k+1}([B, c\gamma]) \approx \frac{\gamma^2\|r\|^2}{1+\gamma^2\|y\|^2} = \|[B, c\gamma] \begin{pmatrix} -y\gamma \\ 1 \end{pmatrix} \|^2 / \| \begin{pmatrix} -y\gamma \\ 1 \end{pmatrix} \|^2;$$

so the STLS distance is well approximated using the Rayleigh quotient corresponding to the unique LS solution of $By\gamma = c\gamma - r\gamma$. This was pointed out by Åke Björck in a personal communication, and may help to explain the behaviour of algorithms proposed in [2]. Alternatively, some bounds here might be rederived via Rayleigh quotient theory.

5 Comparison with other bounds

We now relate our bounds to previous work. Kasenally and Simoncini [10] examined a somewhat related problem for the case of Krylov subspace methods, but did not develop any of the bounds given here. However in a personal communication Simoncini pointed out that if we restrict our discussion to Krylov subspace methods, the equivalent of our weaker lower bound

$\lambda_r \le \|r\|$ in (4.4) can be obtained from their results too. An explanation of this relationship would require a detailed description of the methods analysed in [10], which is beyond the scope of this paper. The best previously published bounds appear to be those of Van Huffel and Vandewalle [9], and we now show how the relevant bounds of that reference, and a new bound, can be derived from (4.2).

Corollary 5.1 *Under the same conditions and assumptions as in Theorem 4.1, with* $\sigma(\gamma) \equiv \sigma_{min}([B, c\gamma])$, $\tau(\gamma) \equiv \gamma\|r\|/\sigma(\gamma)$,

$$0 \le \frac{\gamma^2\|c\|^2 - \sigma(\gamma)^2}{\|B\|^2} \le \tau(\gamma)^2 - 1 \le \frac{\gamma^2\|c\|^2 - \sigma(\gamma)^2}{\sigma_{min}^2(B)}. \tag{5.1}$$

Proof. Remember $\sigma_1 \equiv \sigma_{max}(B) = \|B\|$, $\sigma_k \equiv \sigma_{min}(B)$, $\theta(\gamma) \equiv \sigma(\gamma)/\sigma_1 \le \delta(\gamma) \equiv \sigma(\gamma)/\sigma_k < 1$ from (4.1). Start with (4.2). To obtain the upper bound, we first note from (2.2)–(2.4)

$$\|a\|^2/\sigma_1^2 \le \|y\|^2 \le \|a\|^2/\sigma_k^2, \qquad \|a\|^2 = \|c\|^2 - \|r\|^2, \tag{5.2}$$

giving

$$\begin{aligned}
\frac{\gamma^2\|r\|^2}{\sigma(\gamma)^2} &\le 1 + \frac{\gamma^2\|y\|^2}{1-\delta(\gamma)^2} \le 1 + \frac{\gamma^2(\|c\|^2 - \|r\|^2)}{\sigma_k^2 - \sigma(\gamma)^2} \\
\frac{\gamma^2\|r\|^2}{\sigma(\gamma)^2} \cdot \frac{\sigma_k^2}{\sigma_k^2 - \sigma(\gamma)^2} &\le 1 + \frac{\gamma^2\|c\|^2}{\sigma_k^2 - \sigma(\gamma)^2} \\
\frac{\gamma^2\|r\|^2}{\sigma(\gamma)^2} &\le 1 - \frac{\sigma(\gamma)^2}{\sigma_k^2} + \frac{\gamma^2\|c\|^2}{\sigma_k^2},
\end{aligned}$$

which proves the upper bound in (5.1).

To obtain the lower bound, again combine (4.2) and (5.2) to show

$$\begin{aligned}
\frac{\gamma^2\|r\|^2}{\sigma(\gamma)^2} &\ge 1 + \frac{\gamma^2\|y\|^2}{1-\theta(\gamma)^2} \ge 1 + \frac{\gamma^2(\|c\|^2 - \|r\|^2)}{\sigma_1^2 - \sigma(\gamma)^2} \\
\frac{\gamma^2\|r\|^2}{\sigma(\gamma)^2} \cdot \frac{\sigma_1^2}{\sigma_1^2 - \sigma(\gamma)^2} &\ge 1 + \frac{\gamma^2\|c\|^2}{\sigma_1^2 - \sigma(\gamma)^2} \\
\frac{\gamma^2\|r\|^2}{\sigma(\gamma)^2} &\ge 1 - \frac{\sigma(\gamma)^2}{\sigma_1^2} + \frac{\gamma^2\|c\|^2}{\sigma_1^2}.
\end{aligned}$$

Also $\sigma(\gamma)^2 \le \gamma^2\|c\|^2$, so the rest of (5.1) follows. □

When $\gamma = 1$ the bound $0 \le \tau(\gamma)^2 - 1$ and the upper bound on $\tau(\gamma)^2 - 1$ in (5.1) are (rearrangements of) the equivalents for our situation of (6.34) and (6.35) in [9]. The stronger lower bound seems new. A slightly weaker upper bound was derived in [6, (2.3)]. Experiments show (see Sect. 7) that our bounds in (4.2) can be significantly better than those in (5.1). The relationship of these bounds is, however, intricate. While (5.1) was derived from (4.2), it is not *always* true that the latter is tighter. When $\delta(\gamma) \approx 1$ and $\|r\| \approx \|c\|$,

it is possible for the upper bound in (5.1) to be smaller than that in (4.2). But in this case $\sigma_{min}([B, c\gamma]) \approx \sigma_{min}(B)$, and then the upper bound in (5.1) becomes the trivial $\|r\| \lesssim \|c\|$. Summarizing, when the upper bound in (5.1) is tighter than the upper bound in (4.2), the former becomes trivial and the later is irrelevant. This behaviour is illustrated by our examples in Sect. 7.

The bounds (5.1) and (4.2) differ because the easily available $\|y\|$ in (4.2) was replaced by its upper and lower bounds in (5.2) to obtain (5.1). But there is another reason (4.2) is preferable to (5.1). The latter bounds require knowledge of $\sigma_{min}(B)$, as well as $\sigma_{min}([B, c\gamma])$. Admittedly (4.1) shows we also need these to know $\delta(\gamma)$ exactly, but, assuming that (1.4) holds, we know $\delta(\gamma) < 1$, and is bounded away from 1 always. In fact there are situations where we know $\delta(\gamma) \ll 1$ (we will show practical examples in Sect. 7). Thus (4.2) is not only simpler and often significantly stronger then (5.1), it is more easily applicable.

6 The STLS–LS relationship: equations for $\tau(\gamma)$ and $\delta(\gamma)$

In the previous section we obtained *bounds* for the STLS ratio $\tau(\gamma)$, the LS residual norm $\|r\|$, and the STLS distance $\sigma_{min}([B, c\gamma])$ by using as little additional information as possible. We derived these from the secular equation in Lemma 2.1. Here we look for *exact* relationships. The elegant Theorem 6.9 of Van Huffel and Vandewalle [9] took a different approach than the secular equation, to relate (for $\gamma = 1$) the TLS distance to the LS residual norm $\|r\|$ using full information on the singular values of B and $[c, B]$. They treated the matrix equation $BX \approx C$, but the proof is almost identical to that for (1.1). We give the proof here for completeness (we allow $\gamma \neq 1$), relevance (we state it in terms of the STLS ratio $\tau(\gamma)$) and for the beauty and brevity of their technique.

Theorem 6.1 *Let $n \times k$ B have rank k, $\gamma > 0$, r solve (1.2), and $\sigma_1(\cdot), \sigma_2(\cdot), \ldots$ denote singular values in nonincreasing order, then*

$$(6.1)\quad \tau(\gamma) \equiv \frac{\gamma\|r\|}{\sigma_{k+1}([B, c\gamma])} = \frac{\sigma_1([B, c\gamma])}{\sigma_1(B)} \cdots \frac{\sigma_k([B, c\gamma])}{\sigma_k(B)} \geq 1.$$

Proof. With r and y solving (1.2), the Schur complement of $B^H B$ in $[B, c\gamma]^H [B, c\gamma]$ is, since $B^H r = 0$,

$$\gamma^2 c^H c - \gamma^2 c^H B(B^H B)^{-1} B^H c = \gamma^2 (c^H c - c^H B y) = \gamma^2 c^H r = \gamma^2 \|r\|^2.$$

Using the Schur complement determinant property (2.5) we have

$$\prod_{i=1}^{k+1} \sigma_i^2([B, c\gamma]) = \det([B, c\gamma]^H [B, c\gamma]) = \det(B^H B)\, \gamma^2 \|r\|^2$$

$$= \prod_{i=1}^{k} \sigma_i^2(B)\, \gamma^2 \|r\|^2,$$

giving the equality in (6.1). The bound (see also Lemma 3.2)) holds since the singular values of B interlace those of $[B, c\gamma]$. □

Note the difference of the squares $\tau(\gamma)^2 - 1$ of the left and right hand sides of the inequality (6.1) is bounded in (4.2). The above leads to a new expression and bounds for $\delta(\gamma)$ in (4.1).

Corollary 6.1 *Under the conditions of Theorem 6.1,*

$$\frac{\sigma_{k+1}([B, c\gamma])}{\sigma_1([B, c\gamma])} \;\le\; \frac{\gamma\|r\|}{\|[B, c\gamma]\|} \le \frac{\gamma\|r\|}{\sigma_1([B, c\gamma])} \cdot \frac{\sigma_1(B)}{\sigma_2([B, c\gamma])} \cdots \frac{\sigma_{k-1}(B)}{\sigma_k([B, c\gamma])}$$

$$(6.2) \qquad = \frac{\sigma_{k+1}([B, c\gamma])}{\sigma_k(B)} = \delta(\gamma) \le \frac{\gamma\|r\|}{\sigma_k([B, c\gamma])} \le \frac{\gamma\|r\|}{\sigma_{min}(B)}.$$

Proof. (6.1) shows $\sigma_{k+1}([B, c\gamma]) \le \gamma\|r\|$ so the lowest bound follows. The equalities follow from the equality in (6.1), and the definition of $\delta(\gamma)$ in (4.1). The remaining bounds hold since the singular values of B interlace those of $[B, c\gamma]$. The rightmost bound generalizes (3.3), since the requirements are less restrictive than in (1.4). □

For what it is worth, these give a new expression relating $\delta(\gamma)$ to $\tau(\gamma)$ via ratios of singular values.

Corollary 6.2 *Under the conditions of Theorem 6.1,*

$$(6.3)\ \delta(\gamma) = \tau(\gamma) \left\{ \frac{\sigma_{k+1}([B, c\gamma])}{\sigma_k([B, c\gamma])} \right\} \left\{ \frac{\sigma_1(B)}{\sigma_1([B, c\gamma])} \cdots \frac{\sigma_{k-1}(B)}{\sigma_{k-1}([B, c\gamma])} \right\}.$$

Proof. This follows from the equalities in (6.1) and (6.2). □
Note that the quantities in parentheses $\{\cdot\}$ are each less than or equal to unity since the singular values of B interlace those of $[B, c\gamma]$.

These relationships look very elegant, and the bounds are useful too. In fact for $k = 1$ the tightest upper and lower bounds on $\delta(\gamma)$ in (6.2) become equalities. We see from (6.2) that if $\gamma\|r\|$ is small compared with $\sigma_k([B, c\gamma])$ then $\delta(\gamma) \ll 1$, but if $\gamma\|r\|$ is not small compared with $\|[B, c\gamma]\|$ then $\delta(\gamma)$ cannot be small. If $[B, c\gamma]$ is well-conditioned in the sense that $\sigma_{min}([B, c\gamma])$ is not too much smaller than $\|[B, c\gamma]\|$, then Corollary 6.1 gives us a very good idea of $\delta(\gamma)$.

The computations we have carried out so far, see for example Sect. 7, suggest that the lower bounds in Corollary 6.1 are often very loose, but that the tighter of the following gives (and usually both give) very good upper bounds:

$$(6.4) \qquad \delta(\gamma) \le \min\{1,\ \frac{\gamma\|r\|}{\sigma_k([B, c\gamma])}\} \le \min\{1,\ \frac{\gamma\|r\|}{\sigma_{min}(B)}\}.$$

Some other interesting relationships for the residual norm $\|r\|$ can be found in [11], but they are not motivated by the STLS–LS comparison.

7 The GMRES relationship and numerical experiments

In order to illustrate our theoretical results on numerical experiments we need examples representing various cases. Instead of giving several examples of the matrix B and the right hand side c we will generate sequences of LS-STLS examples as parts of iterative processes.

It is well known that several iterative methods for solving linear algebraic systems form approximate solutions by generating and solving least squares problems at each iteration step. Applying a properly choosen iterative method to a linear system with the matrix A and the right hand side b we get the right hand side c and the sequence of matrices B_k, on which our results can be conveniently and thoroughly illustrated.

But there is also another and much deeper reason for using iterative methods in our experiments. Theorem 4.1 with its corollaries is very useful in the analysis of iterative solutions of nonsingular linear systems. Considerations on bounding the norm of the residual in iterative methods motivated our work which led to the results presented in this paper.

This section very briefly describes the connection to iterative methods and then illustrates our theoretical results with numerical experiments. As an example of an iterative method we will consider the GMRES method [15]. The reader who does not wish to relate our results to iterative methods can simply skip the brief description of GMRES and take the results described later individually (independently for each individual iteration). For a fixed iteration step the displayed results illustrate our theory for some right hand side and some particular (iteration–dependent) matrix.

For a given n by n unsymmetric nonsingular matrix A and n-vector b, we wish to solve $Ax = b$ using the GMRES method. Given an initial approximation x_0 we form the residual $r_0 = b - Ax_0, \rho_0 = \|r_0\|, v_1 = r_0/\rho_0$, and use v_1 to initiate the Arnoldi process. At step k this forms Av_k, orthogonalizes it against $v_1, v_2, \ldots, v_k$, and if the resulting vector is nonzero, normalizes it to give v_{k+1}, giving ideally (in exact arithmetic) $AV_k = V_{k+1}H_{k+1,k}, V_{k+1}^H V_{k+1} = I_{k+1}, V_{k+1} = [v_1, v_2, \ldots, v_{k+1}]$. Here $H_{k+1,k}$ is a $k+1$ by k upper Hessenberg matrix with elements h_{ij} where $h_{j+1,j} \neq 0$, $j = 1, 2, \ldots, k-1$. If at any stage $h_{k+1,k} = 0$ we would stop with $AV_k = V_k H_{k,k}$. Computationally (in finite precision arithmetic) we are unlikely to reach such a k, and we stop when we assess the norm of the residual is small enough.

In general, at each step we take $x_k = x_0 + V_k y_k$ as our approximation to the solution x, which gives the residual $r_k = b - Ax_k = r_0 - AV_k y_k =$

$v_1\rho_0 - V_{k+1}H_{k+1,k}\,y_k$ where y_k solves the linear least squares problem

$$\|r_k\| = \min_{\tilde{y}} \|e_1\rho_0 - H_{k+1,k}\,\tilde{y}\| = \min_{\tilde{y}} \|v_1\rho_0 - AV_k\tilde{y}\| = \rho_0 \min_{\tilde{y}} \|v_1 - AV_k\tilde{y}\|. \tag{7.1}$$

At any iteration step the relative residual norm $\|r_k\|/\rho_0$ can therefore be viewed as the residual norm for the least squares problem with the matrix AV_k and the right hand side v_1. Consider the STLS problem for the matrix AV_k and the right hand side $v_1\gamma$, where we set the value of the scaling parameter $\gamma = 1$. Assume that v_1 is not orthogonal to the left singular vector subspace of AV_k corresponding to $\sigma_{min}(AV_k)$. Then [12, (3.7)] implies that the smallest singular value $\sigma_{k+1}([v_1, AV_k])$ is less than the smallest singular value $\sigma_k(AV_k)$. Consequently, the STLS problem for the matrix AV_k and the right hand side v_1 has a unique solution, with STLS distance $\sigma_{k+1}([v_1, AV_k])$.

In this way GMRES produces sequences of LS and STLS problems where $c = r_0/\rho_0 = (b - Ax_0)/\rho_0 = v_1, \gamma = 1$, and $B = B_k = AV_k$, $y = y_k/\rho_0$, and $r = r_k/\rho_0$ are changing at each step. Please note that for each GMRES iteration we get a new LS, and corresponding STLS, problem. GMRES experiments will therefore allow us to illustrate the variety of situations which were analysed above in this paper.

We could have choosen different right hand sides (for example $c = r_0$) and different values of the scaling parameter γ, but our present choice is simple and sufficient for illustrating our theory. A detailed study of the possible values of γ in relation to the analysis of GMRES will be presented in [13].

In reasonable iterations with B_k increasing in dimension with k, we will usually have $\sigma_{min}(B_k) \to$ constant > 0, while $\sigma_{min}([v_1, B_k])$ eventually becomes zero. Consequently

$$0 \le \delta_k \equiv \sigma_{min}([v_1, B_k])/\sigma_{min}(B_k) \to 0,$$

and from Corollary 4.2

$$0 \le \frac{\|r_k\| - \sigma_{min}([v_1, B_k])\{\rho_0^2 + \|y_k\|^2\}^{\frac{1}{2}}}{\|r_k\|} \equiv \eta_k \le \frac{\delta_k^2}{1 - \delta_k^2} \to 0, \tag{7.2}$$

where for each step k, η_k corresponds to η_r in (4.8). This is a strong "asymptotic" relationship between the minimum residual norm and the minimum singular value of $[v_1, B_k]$.

We will present results of three GMRES experiments, all of them using matrices from the Rutherford-Boeing collection. In all experiments the modified Gram-Schmidt (MGS) orthogonalization was used for computing the

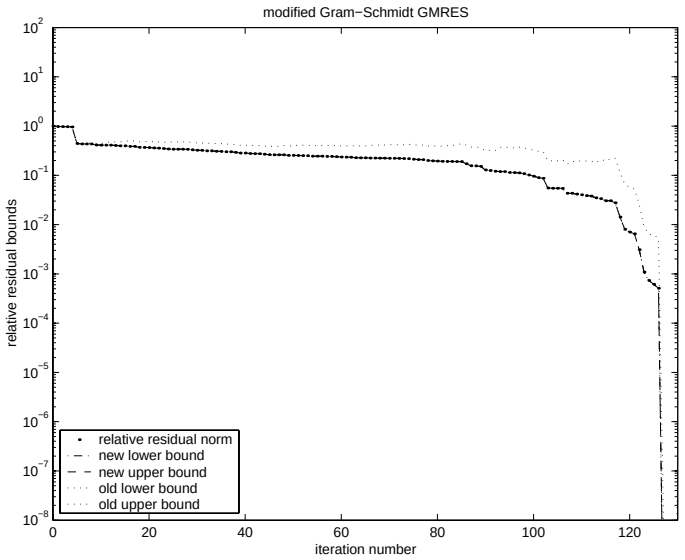

Fig. 7.1. Residual bounds: previous upper and lower bounds from (5.1) (*dotted lines*), then from (4.4) new upper bound (*dashed line*), new lower bound (*dashed-dotted line*), for the relative residual norm (points), computed by MGS GMRES applied to IMPCOLC. Except for the (5.1) upper bound, all values are nearly the same (the corresponding lines coincide) until orthogonality is completely lost (here at the iteration step 128), after which the values are not computed accurately

basis vectors $v_1, v_2, \ldots, v_k$. In the first experiment (using the matrix IMPCOLC) the value of δ_k never becomes close to one and the bounds provided by Theorem 4.1 are very tight. In the second and third experiments (both of them using the matrix WEST0132) δ_k is close to one in some, respectively many, iteration steps. All experiments were performed on a SGI ORIGIN 200 Workstation using MATLAB 6.0, machine precision $\epsilon = 1.11 \times 10^{-16}$.

We first present results of the MGS GMRES algorithm applied to the matrix IMPCOLC, $n = 137$, $\|A\| \approx 120$, $\kappa(A) \approx 1.8 * 10^4$, $b = Ae$, e is the vector of all ones, with $x_0 = \text{randn}(137, 1)$ from MATLAB.

In Fig. 7.1 we plot relative residual bounds, that is, bounds on the relative residual norm $\|b - Ax_k\|/\|r_0\|$, which is denoted by points. The upper and lower bounds from the equivalent of (5.1) (the best previous bounds of Van Huffel and Vandewalle) are denoted by the dotted lines, while the upper bound from (4.4) is given by the dashed line, and the lower bound by the dashed-dotted line. The upper bound from (5.1) is seen to be particularly weak compared with that from (4.4). The lower bound from (5.1) and the bounds from (4.4) almost coincide with the actual values of the relative residual norm.

Figure 7.2 is devoted to the tightness parameters, which show how tight our lower and upper bounds λ_r and μ_r are for this test problem, see

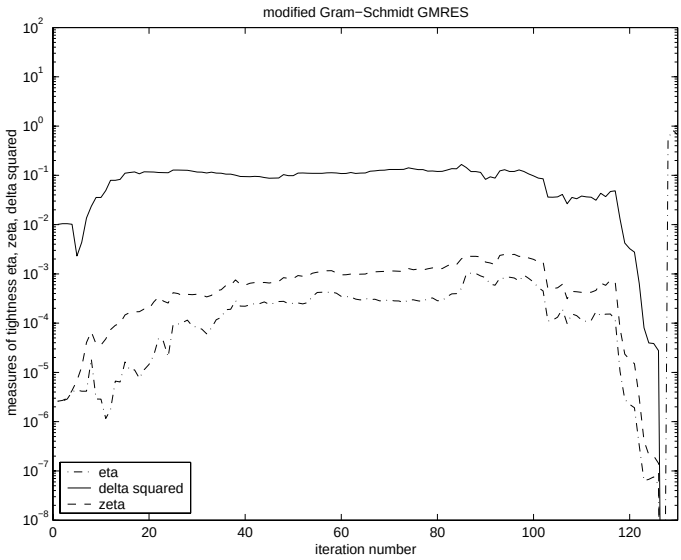

Fig. 7.2. Values of the tightness parameters ${\delta_k}^2$ (*solid line*), η_k (*dashed-dotted line*) and ζ_k (*dashed line*) for MGS GMRES applied to IMPCOLC. The values have little meaning once orthogonality is fully lost

(4.4), (4.8) and (4.9). The solid line shows the values of ${\delta_k}^2$ where $\delta_k \equiv \sigma_{min}([v_1, B_k])/\sigma_{min}(B_k)$, the dashed line the values of ζ_k and the dashed-dotted line the values of η_k for each step k, where δ_k, ζ_k and η_k correspond to $\delta(\gamma)$ in (4.1) and to ζ_r and η_r in (4.8). The values have little meaning once the orthogonality among the basis vectors $v_1, \ldots, v_k$ is fully lost (here at the step $k = 128$), but until that point these computed results follow the theory. For all of the steps δ_k^2 is satisfactory (< 1), while η_k and ζ_k (the relative gaps of the bounds) are quite good (they are reasonably small throughout). But they all decrease impressively when the norm of the relative residual drops towards the machine precision level.

Finally Fig. 7.3 is devoted to bounds on δ_k, which is represented by points. The bounds are those in Corollary 6.1, the weaker upper bound is denoted by the solid line, the stronger by the dashed line, and the tighter lower bound by the dotted line. Note in every case the lower bound is particularly weak, but this does not matter much as we are more interested in upper bounds. The upper bounds are very tight (here δ_k is always significantly less than 1).

The relationship of the new bounds developed in our paper to the best previous bounds is further illustrated by the following two examples. They present results of the MGS GMRES algorithm applied to the matrix WEST0132, $n = 132$, $\|A\| \approx 3.2 * 10^5$, $\kappa(A) \approx 6.4 * 10^{11}$. We will

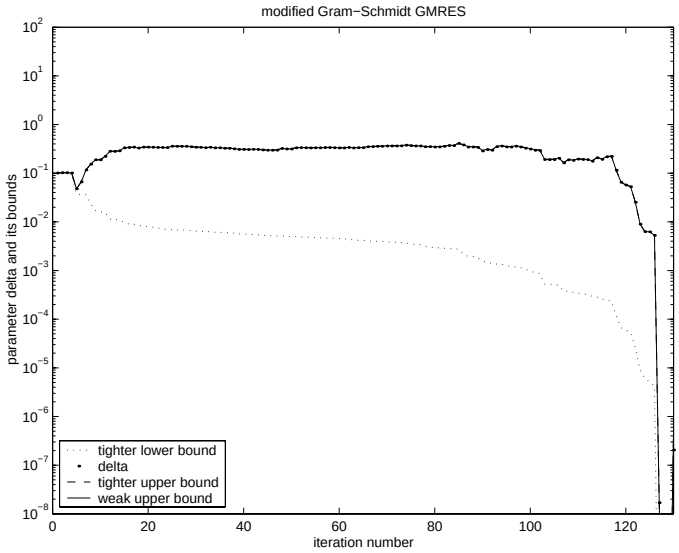

Fig. 7.3. Bounds for δ_k from Corollary 6.1: weak upper bound (*solid line*), tighter upper bound (*dashed line*), δ_k (*points*), all nearly the same; tighter lower bound (*dotted line*), computed by MGS GMRES applied to IMPCOLC

only present the normalized residual bounds with the same meaning and notation as in Fig. 7.1.

Figure 7.4 corresponds to the choice $b = Ae$, with $x_0 = \text{randn}(132, 1)$ from MATLAB. The upper bound from (5.1) is much weaker than the the upper bound from (4.4). For the other bounds all values are nearly the same except for the iterations 18–26 and 122–125 where $\delta_k \to 1$. The maximal value of δ_k during this experiment was 0.99998. Please note that even for $\delta_k \equiv \sigma_{min}([v_1, B_k])/\sigma_{min}(B_k)$ so close to one the new bounds from (4.4) were reasonable, and whenever the upper bound from (4.4) looses its tightness, the upper bound from (5.1) becomes trivial (its value is graphically indistinguishable from one).

Figure 7.5 shows that the upper bound from (5.1) may in some cases be smaller than the upper bound from (4.4). Here we used $b = e$ with $x_0 = 0$ which resulted in a very slow decrease of the relative residual norm. Consequently, when δ_k becomes extremely close to one (in this experiment the maximal value of δ_k was 0.999995) and the upper bound from (4.4) looses its tightness, it may become larger than $\|r_0\|$ and therefore worse then the upper bound from (5.1). The lower bound from (4.4) is (due to the values of δ_k close to one) not tight for most of the iterations, but it is always (and often significantly) better then the lower bound from (5.1). Note also that the tightness of the bounds from (4.4) tends to improve while the tightness of the bounds from (5.1) tends to worsen as the relative residual norm decreases.

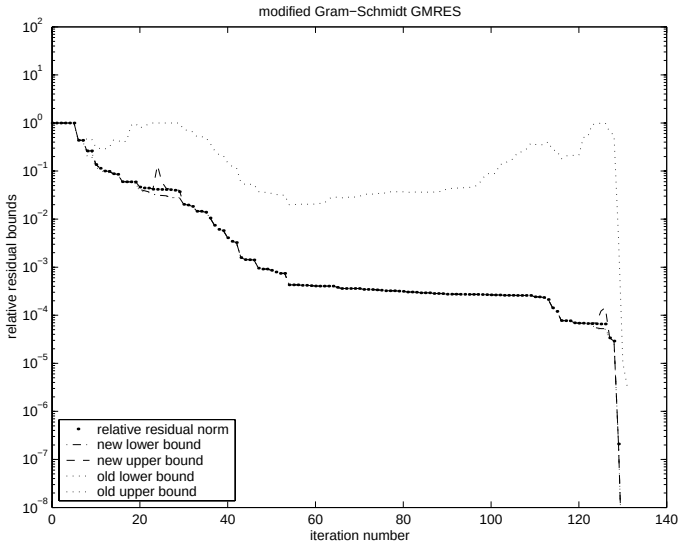

Fig. 7.4. Residual bounds: previous upper and lower bounds from (5.1) (*dotted lines*), then from (4.4) new upper bound (*dashed line*), new lower bound (*dashed-dotted line*), for the relative residual norm (*points*), computed by MGS GMRES applied to WEST0132 with $x = e$ and $x_0 = \text{randn}(132, 1)$. The (5.1) upper bound is weak. For the other bounds all values are nearly the same except for the iterations 18 – 26 and 122 – 125 where $\delta_k \to 1$

8 Summary and conclusion

This paper analysed the relationship between the norm of the residual for the least squares problem for the matrix B and the right hand side c, and the STLS distance $\sigma_{min}([B, c\gamma])$, which is the norm of the corresponding total least squares correction for the TLS problem for the matrix B and the right hand side $c\gamma$. Here γ is a positive parameter which scales the relative sizes of the corrections to B and c. Among other things, we derived new bounds for the LS residual norm $\|r\| = \min_y \|c - By\|$ in terms of the STLS distance $\sigma_{min}([B, c\gamma])$, and proved several important corollaries describing the tightness of the bounds and their dependence on the parameter γ. The bounds were seen to be very good when $\sigma_{min}([B, c\gamma])$ was sufficiently smaller than $\sigma_{min}(B)$. When $\sigma_{min}([B, c\gamma]) \approx \sigma_{min}(B)$, it was shown that the smallest singular value $\sigma_{min}(B)$ and its singular vectors did not play a significant role in the solution of the LS problem. The TLS problem for the matrix B and the right hand side $c\gamma$, $\gamma > 0$ was shown in [12, Sect. 1] to be equivalent to an earlier formulation [14] of the STLS problem for B and c. Our results quantify the relationship between the LS and STLS problems.

We illustrated our theory on the example of the GMRES algorithm which produces sequences of LS and STLS problems. But the relationship between GMRES and the LS and STLS results that has been proven and discussed in

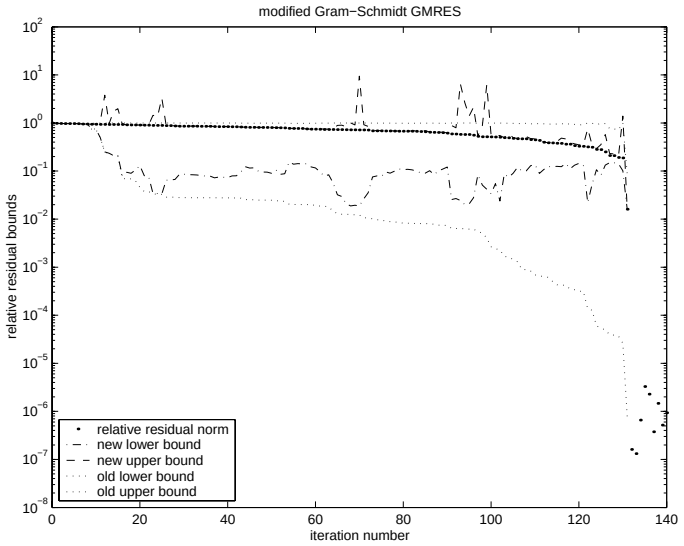

Fig. 7.5. Residual bounds: previous upper and lower bounds from (5.1) (*dotted lines*), then from (4.4) new upper bound (*dashed line*), new lower bound (*dashed-dotted line*), for the normalized residual norm (*points*), computed by MGS GMRES applied to WEST0132 with $b = e$ and $x_0 = 0$. In this extreme situation δ_k is close to one for most of the iterations. Several times it becomes extremely close to one, and the new upper bound from (4.4) is then weaker than the upper bound from (5.1) The new lower bound from (4.4) is always (and for most of the iterations significantly) stronger than the lower bound from (5.1).

this paper is much deeper than shown by this illustration; this relationship is truly fundamental. Our results allow us to explain the role of orthogonality in the finite precision modified Gram-Schmidt GMRES computation and to complete the numerical stability analysis of MGS GMRES started in [6]. In this paper, however, we did not go into the effects of rounding errors (this is why we used experiments in which the loss of orthogonality among the computed Arnoldi basis vectors is minimal for most of the iterations). The subsequent paper [13] will be devoted to the LS–STLS–GMRES relationship and the questions mentioned (but not addressed) here will be treated there in full depth.

Acknowledgements. The authors are indebted to Sabine Van Huffel and Åke Björck for their gracious help and their many valuable suggestions. Comments by Oliver Ernst, Misha Kilmer and Volker Mehrmann were also very helpful.

References

1. Å. Björck: Numerical Methods for Least Squares Problems. Philadelphia PA: SIAM Publications 1996
2. Å. Björck, P. Heggerness, P. Matstoms: Methods for large scale total least squares problems. SIAM J. Matrix Anal. Appl., **22**, 413–429 (2000)
3. G. H. Golub, C. Reinsch: Singular value decomposition and least squares solutions. Numerische Mathematik **14**, 403-420 (1970). Also in "Handbook for Automatic Computation Vol. 2: Linear Algebra", by J. H. Wilkinson, C. Reinsch, (eds.), pp. 134–151. New York: Springer 1971
4. G. H. Golub, C. F. Van Loan: An analysis of the total least squares problem. SIAM J. Numer. Anal. **17**, 883–893 (1980)
5. G. H. Golub, C. F. Van Loan: Matrix Computations. Baltimore MD: The Johns Hopkins University Press, third ed. 1996
6. A. Greenbaum, M. Rozložník, Z. Strakoš: Numerical behavior of the modified Gram-Schmidt GMRES implementation. BIT **37:3**, 706–719 (1997)
7. N. J. Higham: Accuracy and Stability of Numerical Algorithms. Philadelphia PA: SIAM Publications 1996
8. R. A. Horn, C. R. Johnson: Matrix Analysis. Cambridge: Cambridge University Press 1985
9. S. Van Huffel, J. Vandewalle: The Total Least Squares Problem: Computational Aspects and Analysis. Philadelphia PA: SIAM Publications 1991
10. E. M. Kasenally, V. Simoncini: Analysis of a minimum perturbation algorithm for nonsymmetric linear systems. SIAM J. Numer. Anal. **34**, 48–66 (1997)
11. J. Liesen, M. Rozložník, Z. Strakoš: On Convergence and Implementation of Residual Minimizing Krylov Subspace Methods. Submitted to SIAM J. Sci. Comput., September 2000
12. C. C. Paige, Z. Strakoš: Scaled total least squares fundamentals. Accepted for publication in Numerische Mathematik, February 2001
13. C. C. Paige, Z. Strakoš: Residual behaviour in minimum residual Krylov subspace methods. Submitted to SIAM J. Sci. Comput., October 2000
14. B. D. Rao: Unified treatment of LS, TLS and truncated SVD methods using a weighted TLS framework. In: Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modelling, pp. 11–20, S. Van Huffel (ed.), Philadelphia PA: SIAM 1997
15. Y. Saad, M. Schultz, GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Statist. Comput. **7**, 856–869 (1986)