

ESTIMATING THE MINIMAL BACKWARD ERROR IN LSQR

Pavel Jiránek¹

David Tittley-Peloquin²

CERFACS Technical Report TR-PA-09-77

¹CERFACS, 42 Avenue Gaspard Coriolis, Toulouse, France & Faculty of Mechatronics, Technical University of Liberec, Studentská 2, 46117 Liberec, Czech Republic.

E-mail: jirane@cerfacs.fr.

²McGill University, School of Computer Science, 3480 University Street, McConnell Engineering Building, Room 318, Montreal, Quebec, Canada, H3A 2A7.

E-mail: dtittle@cs.mcgill.ca.

Abstract In this paper we propose practical and efficiently-computable stopping criteria for the iterative solution of large sparse linear least squares (LS) problems. Although we focus our discussion on the algorithm LSQR of Paige and Saunders, many ideas discussed here are also applicable to other conjugate gradients type algorithms. We review why the 2-norm of the projection of the residual vector onto the range of A is a useful measure of convergence, and show how this projection can be estimated efficiently at every iteration of LSQR. We also give practical and cheaply-computable estimates of the minimal backward error for the LS problem.

Keywords linear least squares, iterative methods, large sparse matrix problems, stopping criteria, backward perturbation analysis.

Mathematics Subject Classification (2000) 65F10, 65F20, 65F50, 65G50.

The work of the first author was supported by the grant No. 201/09/P464 of the Grant Agency of the Czech Republic. The work of the second author was supported by an NSERC of Canada PGS-D Fellowship.

1 Introduction

Given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, the linear least squares (LS) problem is

$$\min_x \|b - Ax\|_2. \quad (1)$$

It is well known that solving (1) is equivalent to solving the normal equations

$$A^T Ax = A^T b. \quad (2)$$

We assume throughout that A has full column rank. Under this assumption, the unique solution \hat{x} to (1) and (2) is

$$\hat{x} = (A^T A)^{-1} A^T b \equiv A^\dagger b. \quad (3)$$

See for example [2] and [8] for useful background.

In this paper we propose practical stopping criteria for the iterative solution of large sparse LS problems. In particular, we consider the algorithm LSQR of Paige and Saunders [17, 18], which we briefly review in Section 2. Our results, however, could also be applicable to other conjugate gradients type algorithms, such as the method of conjugate gradients [12] applied to the system of normal equations (2), called CGLS in [18]; see also [2, 3].

In Section 3 we define what we mean by an acceptable LS solution and review two conditions recently presented in [5] to determine if a given iterate x_k is an acceptable LS solution. The first criterion requires the computation of the projection $\|P_A r_k\|_2$, where $P_A \equiv AA^\dagger$ is the orthogonal projector onto the range of A and $r_k \equiv b - Ax_k$, while the second involves the computation of the so-called minimal backward error for the LS problem, which we denote μ .

Both $\|P_A r_k\|_2$ and μ are much too expensive to compute to be used directly in large sparse applications. A contribution of this paper is finding estimates of these two quantities that can be computed efficiently at every iteration of LSQR. We discuss methods to estimate the projection $\|P_A r_k\|_2$ in Section 4 and the backward error μ in Section 5.

We generally use upper-case letters for matrices, lower-case Roman letters for vectors and indices, and lower-case Greek letters for scalars. The vector e_j is the j -th column of the unit matrix I . We use $\|A\|_2$ and $\|A\|_F$ to denote the 2-norm and Frobenius norm of A , respectively, and $\|A\|_{2,F}$ when either can be used (consistently throughout an expression). Assuming A has full column rank, its Moore-Penrose generalized inverse is given by $A^\dagger \equiv (A^T A)^{-1} A^T$. We use $P_A \equiv AA^\dagger$ and $P_A^\perp \equiv I - P_A$ for the orthogonal projectors onto the range of A and its orthogonal complement (the null-space of A^T), respectively. Finally, $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ stand for the largest and smallest singular value of A , respectively, and $\kappa_{2,F}(A) \equiv \|A\|_{2,F} \|A^\dagger\|_{2,F}$.

Throughout this paper we illustrate our results with convergence plots produced using MATLAB with IEEE 754 double precision arithmetic with the unit roundoff $u \approx 10^{-16}$. Test problem 1 comes from a surveying application. The matrix A is “well1850.mtx”, a 1850×712 matrix from the Matrix Market [4] with 8755 non-zero entries, 2-norm 1.7, and 2-norm condition number 1.1×10^2 . In test problem 2 we use an $m \times n$ matrix $A = U\Sigma V^T$, where $m = 800$ and $n = 200$. The following factors are used: $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are the full orthogonal Q factors of the QR decomposition of random matrices, and $\Sigma = \text{diag}(\sigma_i) \in \mathbb{R}^{m \times n}$ with

$$\sigma_i \equiv \left(\lfloor (i-1+5)/5 \rfloor \cdot 5/n \right)^3, \quad i = 1, \dots, n.$$

Integer division is used here to obtain repeated singular values. This matrix has 2-norm 1.0 and 2-norm condition number 6.4×10^4 . For each test problem the right-hand side vector b is formed as follows:

$$b = A[n, n-1, \dots, 1]^T + 10^{-5}[1, 2, \dots, m]^T.$$

2 Algorithm LSQR

In this section we give a brief overview of the method LSQR proposed in [17, 18]. The algorithm is based on a variant of the Golub-Kahan iterative bidiagonalization procedure [7, §2], which, given the matrix A and the starting vector b , generates the sequences of vectors u_i, v_i and scalars α_i, β_i satisfying the recurrences

$$\begin{aligned} \beta_1 u_1 &= b, & \alpha_1 v_1 &= A^T u_1, \\ \text{for } k &= 1, 2, \dots \\ \beta_{k+1} u_{k+1} &= Av_k - \alpha_k u_k, \\ \alpha_{k+1} v_{k+1} &= A^T u_{k+1} - \beta_{k+1} v_k. \end{aligned} \quad (4)$$

The coefficients $\alpha_i \geq 0$ and $\beta_i \geq 0$ are computed such that $\|u_i\|_2 = \|v_i\|_2 = 1$. Gathering the vectors $U_k \equiv [u_1, \dots, u_k]$, $V_k \equiv [v_1, \dots, v_k]$, and defining the lower bidiagonal matrices $B_k \in \mathbb{R}^{(k+1) \times k}$ and \bar{B}_k by

$$B_k \equiv \begin{bmatrix} \alpha_1 & & & & \\ \beta_2 & \alpha_2 & & & \\ & \beta_3 & \ddots & & \\ & & \ddots & \alpha_k & \\ & & & \beta_{k+1} & \end{bmatrix}, \quad \bar{B}_k \equiv [B_k, \alpha_{k+1} e_{k+1}], \quad (5)$$

the relations (4) can be written in the compact matrix form

$$U_{k+1}(\beta_1 e_1) = b, \quad (6a)$$

$$AV_k = U_{k+1} B_k, \quad (6b)$$

$$A^T U_{k+1} = V_k B_k^T + \alpha_{k+1} v_{k+1} e_{k+1}^T = V_{k+1} \bar{B}_k^T. \quad (6c)$$

In exact arithmetic the matrices U_{k+1} and V_{k+1} each have orthonormal columns; however, the orthogonality is lost very quickly in presence of rounding errors. The bidiagonalization algorithm can be derived by considering the Lanczos process [15] applied to the symmetric indefinite system

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix},$$

which is the augmented system associated to the LS problem (1), cf. [2, §1.1.4].

The k -th iterate of LSQR has the form $x_k = V_k y_k$, where the vector of coordinates y_k is chosen such that the residual $r_k \equiv b - Ax_k$ has minimal 2-norm. Using (6a) and (6b), r_k can be expressed as

$$r_k = b - AV_k y_k = U_{k+1} t_k, \quad t_k \equiv \beta_1 e_1 - B_k y_k. \quad (7)$$

Therefore, r_k has the minimal 2-norm if and only if y_k is the solution of the LS problem

$$\min_y \|\beta_1 e_1 - B_k y\|_2. \quad (8)$$

It can be shown by induction that

$$\text{range}(V_k) = \text{span} \left\{ A^T b, (A^T A) A^T b, \dots, (A^T A)^{k-1} A^T b \right\} = \mathcal{K}_k(A^T A, A^T b),$$

i.e., the columns of V_k form an orthonormal basis of the Krylov subspace generated by the matrix $A^T A$ and the vector $A^T b$. Thus x_k is the minimum residual approximation of \hat{x} in the Krylov subspace $\mathcal{K}_k(A^T A, A^T b)$. Because y_k is the exact solution of (8), we have using (6b) and (7) that

$$r_k^T AV_k = r_k^T U_{k+1} B_k = (\beta_1 e_1 - B_k y_k)^T B_k = 0. \quad (9)$$

In other words, the LSQR approximation x_k is characterized by

$$x_k \in \mathcal{K}_k(A^T A, A^T b), \quad r_k = b - Ax_k \perp AK_k(A^T A, A^T b). \quad (10)$$

Note that the bidiagonalization algorithm breaks down when either $\alpha_{k+1} = 0$ or $\beta_{k+1} = 0$ in (4). In such a case the iterate x_k is the exact solution of the least squares problem (1), but $\beta_{k+1} = 0$ can happen only if the system $Ax = b$ in (1) is consistent; see for example [16].

The bidiagonal LS problem (8) can be solved by a successive transformation of B_k to upper triangular form using a product of reflections Q_k , leading to

$$Q_k[B_k, \beta_1 e_1] = \begin{bmatrix} R_k & f_k \\ 0 & \bar{\phi}_{k+1} \end{bmatrix}, \quad R_k \equiv \begin{bmatrix} \rho_1 & \theta_2 & & \\ & \rho_2 & \ddots & \\ & & \ddots & \theta_k \\ & & & \rho_k \end{bmatrix}, \quad f_k \equiv \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_k \end{bmatrix}. \quad (11)$$

The k -th reflection is designed to zero β_{k+1} with the element directly above it, and has the form

$$\begin{bmatrix} c_k & s_k \\ s_k & -c_k \end{bmatrix} \begin{bmatrix} \bar{\rho}_k & 0 & \bar{\phi}_k \\ \beta_{k+1} & \alpha_{k+1} & 0 \end{bmatrix} = \begin{bmatrix} \rho_k & \theta_{k+1} & \phi_k \\ 0 & \bar{\rho}_{k+1} & \bar{\phi}_{k+1} \end{bmatrix}, \quad (12)$$

where $\bar{\rho}_1 = \alpha_1$ and $\bar{\phi}_1 = \beta_1$.

From (11), the solution of the bidiagonal LS problem (8) is $y_k = R_k^{-1} f_k$, while the minimal residual t_k and corresponding residual $r_k = b - Ax_k$ in (7) can be expressed, respectively, as

$$t_k = \bar{\phi}_{k+1} Q_k^T e_{k+1}, \quad r_k = U_{k+1} t_k = \bar{\phi}_{k+1} U_{k+1} Q_k^T e_{k+1}. \quad (13)$$

Note from the above that $\|r_k\|_2 = \bar{\phi}_{k+1}$, so the 2-norm of the residual vector can easily be obtained without actually computing the residual vector r_k . The relation $\|A^T r_k\|_2 = \hat{\phi}_{k+1} \alpha_{k+1} |c_k|$ can also easily be derived, while $\|A\|_2$ is usually well estimated by $\|B_k\|_2$; see [18, §5].

A short recurrence for the x_k can be obtained by noticing that the first $k-1$ elements of f_k in (11) form the vector f_{k-1} from the previous step. Defining

$$W_k \equiv [w_1, \dots, w_k] \equiv V_k R_k^{-1} D_k, \quad D_k \equiv \text{diag}(R_k), \quad (14)$$

we obtain

$$x_k = V_k y_k = W_k D_k^{-1} f_k = [W_{k-1}, w_k] \begin{bmatrix} D_{k-1}^{-1} & \\ & \rho_k^{-1} \end{bmatrix} \begin{bmatrix} f_{k-1} \\ \phi_k \end{bmatrix} = x_{k-1} + \frac{\phi_k}{\rho_k} w_k. \quad (15)$$

A short recurrence for the w_k can also be obtained. Solving $W_{k+1} D_{k+1}^{-1} R_{k+1} = V_{k+1}$ by forward substitution gives

$$w_{k+1} = v_{k+1} - \frac{\theta_{k+1}}{\rho_k} w_k, \quad w_1 = v_1.$$

The implementation of LSQR shown in Algorithm 1 requires two matrix-vector multiplications with A and A^T per iteration and stores only the latest columns of U_{k+1} , V_{k+1} , and W_{k+1} together with the actual iterate x_k .

3 Acceptable LS Solutions

Following the nomenclature in [5], we say that an iterate $x_k \in \mathbb{R}^n$ an *acceptable LS solution* when it is the exact solution of a LS problem within some specified range of relative errors in the data. In other words, an iterate x_k is an acceptable LS solution if and only if there exist perturbations E and f satisfying

$$(A + E)^T (A + E) x_k = (A + E)^T (b + f), \quad \|E\|_{2,F} \leq \alpha \|A\|_{2,F}, \quad \|f\|_2 \leq \beta \|b\|_2, \quad (16)$$

Algorithm 1 LSQR

- Initialize:
 $\beta_1 u_1 = b$, $\alpha_1 v_1 = A^T u_1$, $w_1 = v_1$, $x_0 = 0$, $\bar{\phi}_1 = \beta_1$, $\bar{\rho}_1 = \alpha_1$
- Main loop:
for $k = 1, 2, 3, \dots$
 - Continue the bidiagonalization:
 $\beta_{k+1} u_{k+1} = A v_k - \alpha_k u_k$
 $\alpha_{k+1} v_{k+1} = A^T u_{k+1} - \beta_{k+1} v_k$
 - Form the k -th reflection:
 $\rho_k = (\bar{\rho}_k^2 + \beta_{k+1}^2)^{1/2}$, $c_k = \bar{\rho}_k / \rho_k$, $s_k = \beta_{k+1} / \rho_k$
 - Apply the reflection:
 $\theta_{k+1} = s_k \alpha_{k+1}$, $\bar{\rho}_{k+1} = -c_k \alpha_{k+1}$, $\phi_k = c_k \bar{\phi}_k$, $\bar{\phi}_{k+1} = s_k \bar{\phi}_k$
 - Update x_k and w_{k+1} :
 $x_k = x_{k-1} + (\phi_k / \rho_k) w_k$
 $w_{k+1} = v_{k+1} - (\theta_{k+1} / \rho_k) w_k$
 - Test for convergence and exit if x_k is an acceptable LS solution.

for some chosen values of α and β (distinct from the elements α_k and β_k of B_k).

One choice for the parameters α and β of particular interest is $\alpha = \mathcal{O}(u)$ and $\beta = \mathcal{O}(u)$, where u denotes the unit roundoff. If (16) holds with this choice of parameters then x_k is a so-called *backward stable LS solution*. Statistical arguments can also be used to pick α and β . For example, one can show that the statistical stopping criteria proposed by Arioli and Gratton in [1, §3.1.1] are triggered if and only if (16) holds with $\alpha = 0$ and β chosen using the inverse cumulative distribution function of the chi-squared probability distribution. We refer the interested reader to the discussion in [1].

In the following we review two conditions that can be used to verify if (16) holds, one of which is both necessary and sufficient.

For any chosen α and β , (16) holds if and only if $\xi_{2,F}(x_k, \alpha, \beta) \leq 1$, where

$$\xi_{2,F}(x_k, \alpha, \beta) \equiv \min_{E,f,\xi} \left\{ \xi : (A + E)^T [b + f - (A + E)x_k] = 0, \right. \\ \left. \|E\|_{2,F} \leq \xi \alpha \|A\|_{2,F}, \|f\|_2 \leq \xi \beta \|b\|_2 \right\}. \quad (17)$$

The analytical solution of (17) remains an open question. Below we present tight bounds on $\xi_{2,F}(x_k, \alpha, \beta)$.

Stopping criteria for the iterative solution of large sparse LS problems are commonly based on upper bounds on $\xi_{2,F}(x_k, \alpha, \beta)$. For example, it is recommended in [18, §5] to stop as soon as an iterate x_k with corresponding residual $r_k \equiv b - Ax_k$ satisfy

$$\|r_k\|_2 \leq \alpha \|A\|_{2,F} \|x_k\|_2 + \beta \|b\|_2 \quad \text{or} \quad \|A^T r_k\|_2 \leq \alpha \|A\|_{2,F} \|r_k\|_2. \quad (18)$$

These stopping criteria are based on the following upper bounds on $\xi_{2,F}(x_k, \alpha, \beta)$ of Rigal and Gaches [19] and Stewart [20], [21, Theorem 5.5], respectively:

$$\xi_{2,F}(x_k, \alpha, \beta) \leq \frac{\|r_k\|_2}{\alpha \|A\|_{2,F} \|x_k\|_2 + \beta \|b\|_2}, \\ \xi_{2,F}(x_k, \alpha, \beta) \leq \frac{\|A^T r_k\|_2}{\alpha \|A\|_{2,F} \|r_k\|_2}.$$

They give sufficient but not necessary conditions for $\xi_{2,F}(x_k, \alpha, \beta) \leq 1$. It was shown in [5, §4] that in many practical situations the stopping criteria in (18) can detect an acceptable LS solution several iterations too late, or even fail altogether to detect that an acceptable LS solution has been obtained.

The following asymptotically tight upper bound on $\xi_{2,F}(x_k, \alpha, \beta)$ was given in [5, Theorem 5.1].

Lemma 3.1. *Given $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $x_k \in \mathbb{R}^n$, define $r_k \equiv b - Ax_k$ and $\xi_{2,F}(x_k, \alpha, \beta)$ as in (17). Then*

$$\xi_{2,F}(x_k, \alpha, \beta) \leq \psi_{2,F}(x_k, \alpha, \beta) \equiv \frac{\|PAr_k\|_2}{\alpha\|A\|_{2,F}\|x_k\|_2 + \beta\|b\|_2}, \quad (19)$$

and letting \hat{x} denote the true LS solution of (1),

$$\lim_{x_k \rightarrow \hat{x}} \frac{\xi_{2,F}(x_k, \alpha, \beta)}{\psi_{2,F}(x_k, \alpha, \beta)} = 1. \quad (20)$$

Recall that a given vector $x_k \in \mathbb{R}^n$ is an acceptable LS solution if and only if $\xi_{2,F}(x_k, \alpha, \beta) \leq 1$. As a result of Lemma 3.1, if

$$\|PAr_k\|_2 \leq \alpha\|A\|_{2,F}\|x_k\|_2 + \beta\|b\|_2 \quad (21)$$

then $\xi_{2,F}(x_k, \alpha, \beta) \leq 1$ and x_k is an acceptable LS solution. From (20) we can expect the bound in (19) to be tight if x_k is sufficiently close to the true LS solution \hat{x} . Numerical tests performed in [5] suggest that this bound is often tight even when x_k is very far from \hat{x} .

The minimization problem in Lemma 3.2 was solved in [25, §2]. It is commonly referred to as a *minimal backward error problem*; see for example [9] and the references therein.

Lemma 3.2. *Given $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $0 \neq x_k \in \mathbb{R}^n$ and $\theta > 0$, define*

$$r_k \equiv b - Ax_k, \quad \omega_k \equiv \frac{\theta\|r_k\|_2}{\sqrt{1 + \theta^2\|x_k\|_2^2}}, \quad N_k \equiv [A, \omega_k(I - r_k r_k^\dagger)].$$

Then

$$\begin{aligned} \mu(x_k, \theta) &\equiv \min_{\Delta A, \Delta b} \left\{ \|\Delta A, \theta \Delta b\|_F : (A + \Delta A)^T [(b + \Delta b) - (A + \Delta A)x_k] = 0 \right\} \\ &= \min \left\{ \omega_k, \sigma_{\min}(N_k) \right\}. \end{aligned} \quad (22)$$

We use the notation

$$\mu(x_k, \infty) \equiv \lim_{\theta \rightarrow \infty} \mu(x_k, \theta) = \min_{\Delta A} \left\{ \|\Delta A\|_F : (A + \Delta A)^T [b - (A + \Delta A)x_k] = 0 \right\}.$$

Gu [10] gave a relationship between the quantities $\mu(x_k, \infty)$ and $\xi_{2,F}(x_k, \alpha, \beta)$ in the special case when $\alpha = \beta$. The following lemma, proven in [5, Theorem 6.2], shows how we can use $\mu(x_k, \theta)$ with an appropriate finite θ in (22) to determine if x_k is an acceptable LS solution in the Frobenius norm for any choice of α and β in (17).

Lemma 3.3. *Given full column rank $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $0 \neq x_k \in \mathbb{R}^n$, define $r_k \equiv b - Ax_k$, $\xi_F(x_k, \alpha, \beta)$ as in (17) and $\mu(x_k, \theta)$ as in (22). Let*

$$\theta = \hat{\theta} \equiv \frac{\alpha\|A\|_F}{\beta\|b\|_2}. \quad (23)$$

Then

$$\xi_F(x_k, \alpha, \beta) \leq \frac{\mu(x_k, \hat{\theta})}{\alpha \|A\|_F} \leq \sqrt{2} \xi_F(x_k, \alpha, \beta). \quad (24)$$

Recall that x_k is an acceptable LS solution if and only if $\xi_{2,F}(x_k, \alpha, \beta) \leq 1$, but that at present no explicit formula is known for computing $\xi_{2,F}(x_k, \alpha, \beta)$. Lemma 3.3 states that the quantity $\mu(x_k, \hat{\theta})/(\alpha \|A\|_F)$ always lies within a factor $\sqrt{2}$ of $\xi_F(x_k, \alpha, \beta)$. The following is therefore a near optimal test to verify if a given iterate x_k is an acceptable LS solution in the Frobenius norm, for any choice of α and β in (17):

$$\mu(x_k, \hat{\theta}) \leq \alpha \|A\|_F, \quad \text{with } \hat{\theta} = \frac{\alpha \|A\|_F}{\beta \|b\|_2}. \quad (25)$$

In summary, the quantities $\|P_{Ar_k}\|_2$ and $\mu(x_k, \hat{\theta})$ can be used to determine if an iterate x_k is an acceptable LS solution or a backward stable LS solution. Both these quantities, of course, are much too expensive to compute to be used directly in stopping criteria for the iterative solution of practical large sparse problems. In the next sections we present new bounds on and estimates of $\|P_{Ar_k}\|_2$ and $\mu(x_k, \theta)$ that can be estimated efficiently at every iteration of LSQR. These estimates can be used in (21) and (25), respectively, to give practical stopping criteria for the iterative solution of large sparse LS problems.

4 Estimating $\|P_{Ar_k}\|_2$ efficiently

In this section we discuss ways to estimate $\|P_{Ar_k}\|_2$ efficiently in LSQR.

The following lemma relates the quantities $\|P_{Ar_k}\|_2$ and $\|A^T r_k\|_2$. It can easily be proven using the singular value decomposition of A ; see for example [5].

Lemma 4.1. *Given $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $x_k \in \mathbb{R}^n$, define $r_k \equiv b - Ax_k$. Then*

$$\|A^T r_k\|_2 = \bar{\sigma}_k \|P_{Ar_k}\|_2$$

for some $\bar{\sigma}_k$ in the closed interval $[\sigma_{\min}(A), \sigma_{\max}(A)]$.

The following bounds on $\|P_{Ar_k}\|_2$ follow immediately from Lemma 4.1:

$$\frac{\|A^T r_k\|_2}{\sigma_{\max}(A)} \leq \|P_{Ar_k}\|_2 \leq \frac{\|A^T r_k\|_2}{\sigma_{\min}(A)}. \quad (26)$$

The above lower bound can be estimated at virtually no extra cost in LSQR, since reliable estimates of $\|A^T r_k\|_2$ and $\sigma_{\max}(A)$ are available essentially free at every iteration; see Section 2. Accurately estimating the smallest singular value of A , however, is much more challenging, so the above upper bound is not as practically useful.

These bounds are plotted in Figure 1. Here we have explicitly computed $\|A^T r_k\|_2$ and the extreme singular values of A . Using LSQR's estimates of $\|A^T r_k\|_2$ and $\sigma_{\max}(A)$ in the lower bound gave very similar results. It is easy to see from (26) that the bounds are simultaneously tight only in very well-conditioned problems (in which $\kappa_2(A) = \sigma_{\max}(A)/\sigma_{\min}(A) \approx 1$). In the problems we have tested, we have observed that the lower bound is usually tighter than the upper bound, especially late in the LSQR iteration process.

With $\hat{x} = A^\dagger b$ denoting the true LS solution of (1), the quantity $\|\hat{x} - x_k\|_{A^T A} \equiv \|A(\hat{x} - x_k)\|_2$ is often referred to as the *energy norm* of the error. The following lemma shows that the projection $\|P_{Ar_k}\|_2$ is in fact $\|\hat{x} - x_k\|_{A^T A}$.

Lemma 4.2. *Given $A \in \mathbb{R}^{m \times n}$ with full column rank and $b \in \mathbb{R}^m$, let \hat{x} be the true LS solution of (1). For any approximate solution $x_k \in \mathbb{R}^n$, define $r_k \equiv b - Ax_k$. Then*

$$\|P_{Ar_k}\|_2 = \|\hat{x} - x_k\|_{A^T A}. \quad (27)$$

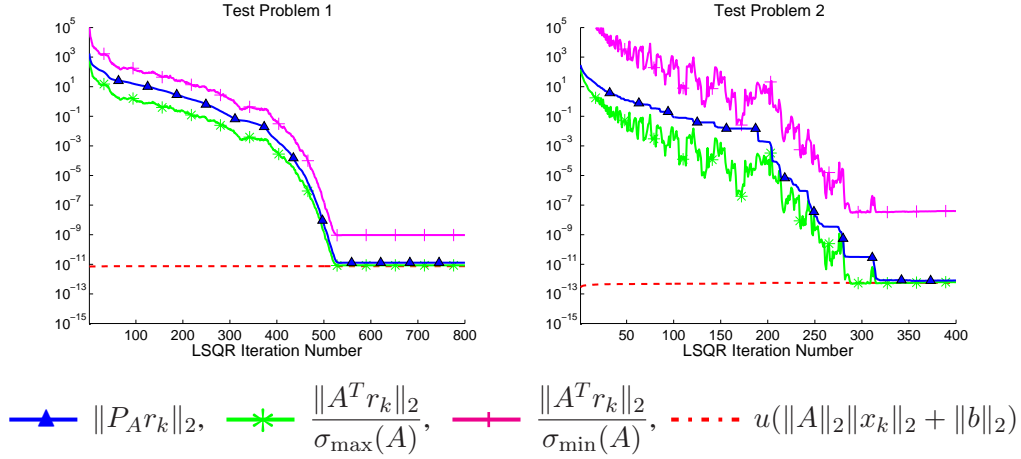


Figure 1: The bounds (26) on $\|P_A r_k\|_2$

Proof. Since $P_A = AA^\dagger$ and the true LS solution is $\hat{x} = A^\dagger b$,

$$\|P_A r_k\|_2 = \|AA^\dagger(b - Ax_k)\|_2 = \|A(\hat{x} - x_k)\|_2 = \|\hat{x} - x_k\|_{A^T A}.$$

□

It follows from (10) that, at least in exact arithmetic, LSQR is equivalent to the method of conjugate gradients (CG) of Hestenes and Stiefel [12] applied to the normal equations (2). Many estimates of the energy norm of the error in CG have been studied, as discussed for example in [22, 23], and these can be extended to estimate $\|P_A r_k\|_2$ in LSQR. In the following we derive one such estimate, originally proposed for CG in [12, Theorem 6.1], and recently extended to CGLS in [1], directly for LSQR. We also provide a useful new way to interpret this estimate.

Theorem 4.1. *Given $A \in \mathbb{R}^{m \times n}$ with full column rank and $b \in \mathbb{R}^m$, let \hat{x} be the true LS solution of (1) and x_k be the k -th iterate of LSQR with $r_k \equiv b - Ax_k$. Then using the notation of Algorithm 1, for any non-negative scalars k and d ,*

$$\|\hat{x} - x_k\|_{A^T A}^2 = \|\hat{x} - x_{k+d}\|_{A^T A}^2 + \sum_{i=k+1}^{k+d} \phi_i^2. \quad (28)$$

Proof. We can write

$$\begin{aligned} \|\hat{x} - x_k\|_{A^T A}^2 &= \|\hat{x} - x_{k+1} + x_{k+1} - x_k\|_{A^T A}^2 \\ &= \|\hat{x} - x_{k+1}\|_{A^T A}^2 + \|x_{k+1} - x_k\|_{A^T A}^2 + 2(r_{k+1} - \hat{r})^T A(x_{k+1} - x_k). \end{aligned} \quad (29)$$

The last term in the above equation is identically 0, since both $\hat{r}^T A = 0$ and from (9)

$$r_{k+1}^T A(x_{k+1} - x_k) = r_{k+1}^T A V_{k+1} \left(y_{k+1} - \begin{bmatrix} y_k \\ 0 \end{bmatrix} \right) = 0.$$

Furthermore in LSQR we have $x_{k+1} = x_k + (\phi_{k+1}/\rho_{k+1})w_{k+1}$; see (15). Thus (29) becomes

$$\|\hat{x} - x_k\|_{A^T A}^2 = \|\hat{x} - x_{k+1}\|_{A^T A}^2 + (\phi_{k+1}/\rho_{k+1})^2 \|Aw_{k+1}\|_2^2.$$

Repeating this argument $d - 1$ times gives

$$\|\hat{x} - x_k\|_{A^T A}^2 = \|\hat{x} - x_{k+d}\|_{A^T A}^2 + \sum_{i=k+1}^{k+d} (\phi_i/\rho_i)^2 \|Aw_i\|_2^2.$$

It remains to show that $\|Aw_i\|_2 = \rho_i$. Using (14) along with (6b) and (11), we obtain

$$Aw_i = AW_i e_i = AV_i R_i^{-1} D_i e_i = U_{i+1} B_i R_i^{-1} D_i e_i = U_{i+1} Q_i^T \begin{bmatrix} I_i \\ 0 \end{bmatrix} \rho_i e_i,$$

and thus $\|Aw_i\|_2 = \rho_i$, completing the proof. \square

We can use Theorem 4.1 to estimate $\|P_{Ar_k}\|_2 = \|\hat{x} - x_k\|_{A^T A}$ as follows. Ignoring the $\|\hat{x} - x_{k+d}\|_{A^T A}^2$ term in (28) gives

$$\|\hat{x} - x_k\|_{A^T A}^2 \geq \sum_{i=k+1}^{k+d} \phi_i^2 \equiv \lambda_d^2(x_k). \quad (30)$$

The hope is that

$$\|\hat{x} - x_k\|_{A^T A}^2 \gg \|\hat{x} - x_{k+d}\|_{A^T A}^2 \quad (31)$$

in (28), and thus that the bound in (30) is tight. This is certainly reasonable for large enough d , since in LSQR the quantity $\|P_{Ar_k}\|_2 = \|\hat{x} - x_k\|_{A^T A}$ is theoretically strictly monotonically decreasing with k and converges to 0; see for example the discussion in [5].

Notice that the estimate $\lambda_d(x_k)$ can be computed at essentially no extra cost in LSQR, but d additional iterations have to be performed in order to estimate $\|\hat{x} - x_k\|_{A^T A}$. A larger value of d results in a longer such delay; however, it also produces a larger difference in (31) and thus a tighter bound in (30). Values of d as small as $d = 5$ can give excellent results in well-conditioned problems, as illustrated with test problem 1 in Figure 2. Higher values of d are required in more ill-conditioned problems, when $\|P_{Ar_k}\|_2$ decreases very slowly or in a staircase pattern and (31) is thus not valid for small d ; see for example test problem 2 in Figure 2. The use of smaller values of d , however, is usually sufficient when a good preconditioner is applied. A better understanding of why this projection can decrease in such a pattern would allow us to make a more informed choice of d , even perhaps to modify d at each iteration. We leave this for a future investigation.

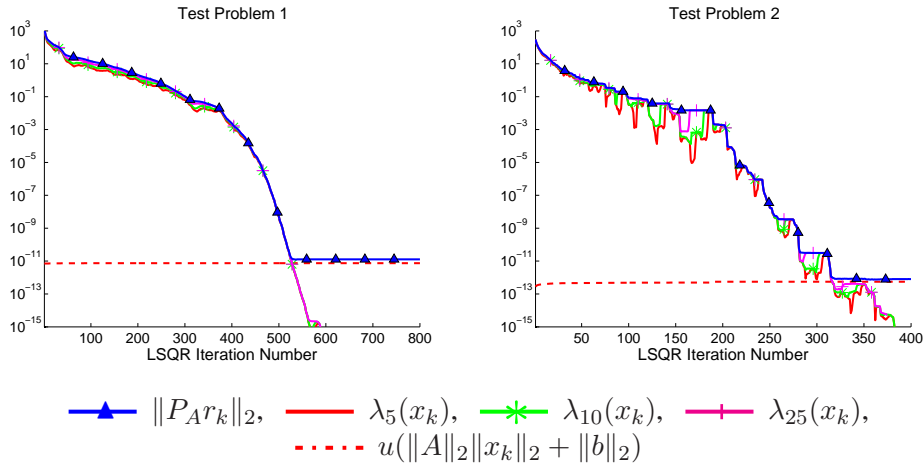


Figure 2: Estimating $\|P_{Ar_k}\|_2$ using $\lambda_d(x_k)$ in (30) with various values of d .

Note in Figure 2 that $\lambda_d(x_k)$ continues to decrease even when $\|P_{Ar_k}\|_2$ hits a plateau near the tolerance $u(\|A\|_2 \|x_k\|_2 + \|b\|_2)$. In practice, as soon as $\|P_{Ar_k}\|_2 = \mathcal{O}(u(\|A\|_2 \|x_k\|_2 + \|b\|_2))$ we should conclude from Lemma 3.1 that x_k is a backward stable iterate and stop the iteration. See also [3] for the discussion on the attainable accuracy of conjugate gradients type methods for LS problems in terms of the backward error.

The following theorem gives another useful way of interpreting the estimate $\lambda_d(x_k)$.

Theorem 4.2. *The estimate $\lambda_d(x_k)$ in (30) satisfies*

$$\lambda_d(x_k) = \|B_{k+d}B_{k+d}^\dagger\hat{t}_k\|_2 \quad (32)$$

where B_{k+d} is the matrix defined in (5) at the iteration step $k+d$ and $\hat{t}_k \equiv [t_k^T, 0]^T$, with t_k given in (7).

Proof. Due to (11), $B_{k+d}^\dagger = [R_{k+d}^{-1}, 0]Q_{k+d}$. Note that

$$Q_{k+d} = H_{k+d} \dots H_1, \quad \hat{Q}_k \equiv \text{diag}(Q_k, I_d) = H_k \dots H_1,$$

where each $H_i \in \mathbb{R}^{(k+d+1) \times (k+d+1)}$ is the elementary reflection described by (12). Using (13), $\hat{t}_k = \hat{Q}_k^T \bar{\phi}_{k+1} e_{k+1}$, where e_{k+1} is the $(k+1)$ -st column of I_{k+d+1} . It follows that

$$\begin{aligned} Q_{k+d}\hat{t}_k &= Q_{k+d}\hat{Q}_k^T \bar{\phi}_{k+1} e_{k+1} = H_{k+d} \dots H_{k+1} \bar{\phi}_{k+1} e_{k+1} \\ &= [0, \dots, 0, \phi_{k+1}, \dots, \phi_{k+d}, \bar{\phi}_{k+d+1}]^T, \end{aligned} \quad (33)$$

and therefore

$$\begin{aligned} B_{k+d}B_{k+d}^\dagger\hat{t}_k &= Q_{k+d}^T \begin{bmatrix} R_{k+d} \\ 0 \end{bmatrix} [R_{k+d}^{-1}, 0]Q_{k+d}\hat{t}_k \\ &= Q_{k+d}^T [0, \dots, 0, \phi_{k+1}, \dots, \phi_{k+d}, 0]^T. \end{aligned}$$

Taking the norm we obtain finally

$$\|B_{k+d}B_{k+d}^\dagger\hat{t}_k\|_2^2 = \sum_{i=k+1}^{k+d} \phi_i^2 = \lambda_d^2(x_k).$$

□

Consider the iterate $x_k = V_k y_k$ with residual r_k computed at step k of LSQR. In the bases V_{k+d} and U_{k+d+1} , the vectors x_k and r_k have coordinates

$$x_k = V_k y_k = V_{k+d} \begin{bmatrix} y_k \\ 0 \end{bmatrix} \equiv V_{k+d} \hat{y}_k, \quad r_k = U_{k+1} t_k = U_{k+d+1} \begin{bmatrix} t_k \\ 0 \end{bmatrix} = U_{k+d+1} \hat{t}_k.$$

Theorem 4.2 states that the estimate $\lambda_d(x_k)$ is precisely the 2-norm of the projected residual $\hat{t}_k = \beta_1 e_1 - B_{k+d} \hat{y}_k$ associated to the projected problem (8) solved by LSQR at the iteration $k+d$.

The estimate $\lambda_d(x_k)$ of $\|P_A r_k\|_2$ allows us to use (21) as a practical stopping criterion in LSQR. In the next section we turn our attention to estimating $\mu(x_k, \hat{\theta})$ in (22) efficiently, using the ideas of Theorem 4.2, so that (25) can also be used as a stopping criterion in practical large sparse applications.

5 Estimating $\mu(x_k, \theta)$

Recall from Lemma 3.3 that we are interested in estimating $\mu(x_k, \hat{\theta})$, with a specific finite $\hat{\theta}$. Many estimates of $\mu(x_k, \infty)$ exist in the literature; see for example [9, §4] and the references therein. These can readily be generalized to estimate $\mu(x_k, \theta)$ for any finite $\theta > 0$. Very few of these estimates, however, are truly suitable for use in large sparse applications, since most either cost $\mathcal{O}(mn^2)$ flops to compute, assume that a factorization of A is available, or are themselves iterative. (Here we are not interested in nesting an iterative method to compute $\mu(x_k, \hat{\theta})$ at each iteration of LSQR.)

In Section 5.1 we give practical new bounds on $\mu(x_k, \theta)$ that are analogous to the estimate $\lambda_d(x_k)$ for $\|P_A r_k\|_2$ given in Section 4. In Section 5.2 we show how an asymptotic estimate of $\mu(x_k, \theta)$ can be estimated efficiently, using the ideas of Theorem 4.2.

5.1 Practical bounds on $\mu(x_k, \theta)$

Suppose that P and Q are orthogonal projectors onto the subspaces $\text{range}(P) \subset \mathbb{R}^m$ and $\text{range}(Q) \subset \mathbb{R}^n$, and consider the least squares problem

$$\min_{z \in \text{range}(Q)} \|P(b - Az)\|_2. \quad (34)$$

Let the columns of the matrices U and V form orthonormal bases of $\text{range}(P)$ and $\text{range}(Q)$, respectively, so that $P = UU^T$ and $Q = VV^T$. Denoting $B = U^T A V$, $c = U^T b$, and $z = V y$, we have

$$\min_{z \in \text{range}(Q)} \|P(b - Az)\|_2 = \min_y \|c - B y\|_2. \quad (35)$$

Note from (6) that choosing $U = U_{k+d+1}$ and $V = V_{k+d}$ leads to $c = \beta_1 e_1$ and $B = B_{k+d}$, that is, the bidiagonal least squares problem (8) solved by LSQR at the iteration step $k + d$, while choosing $U = U_{k+d+1}$ and $V = V_{k+d+1}$ leads to $c = \beta_1 e_1$ and $B = \overline{B}_{k+d}$.

Given an approximate solution of the form $x_k = V \tilde{y}$, here the k -th iterate of LSQR, we can ask how well the vector of coordinates \tilde{y} satisfies the suitably projected least squares problem (35). In other words, we can approximate the value of $\mu(x_k, \theta)$ by computing the minimal backward error defined in Lemma 3.2 of the vector \tilde{y} associated to the projected problem. In the following, we show that for any $d \geq 0$ the choice $U = U_{k+d+1}$ and $V = V_{k+d}$ leads to a lower bound on $\mu(x_k, \theta)$, while $U = U_{k+d+1}$ and $V = V_{k+d+1}$ leads to an upper bound. In the bases V_{k+d} and V_{k+d+1} , the vector x_k has coordinates

$$x_k = V_{k+d} \begin{bmatrix} y_k \\ 0_d \end{bmatrix} = V_{k+d} \hat{y}_k, \quad x_k = V_{k+d+1} \begin{bmatrix} y_k \\ 0_{d+1} \end{bmatrix} \equiv V_{k+d+1} \bar{y}_k.$$

We define

$$\begin{aligned} \underline{\mu}_d(x_k, \theta) \equiv \min_{\Delta B, \Delta c} \left\{ \|\Delta B, \theta \Delta c\|_F : \right. \\ \left. (B_{k+d} + \Delta B)^T [(\beta_1 e_1 + \Delta c) - (B_{k+d} + \Delta B) \hat{y}_k] = 0 \right\}, \end{aligned} \quad (36a)$$

$$\begin{aligned} \bar{\mu}_d(x_k, \theta) \equiv \min_{\Delta B, \Delta c} \left\{ \|\Delta B, \theta \Delta c\|_F : \right. \\ \left. (\overline{B}_{k+d} + \Delta B)^T [(\beta_1 e_1 + \Delta c) - (\overline{B}_{k+d} + \Delta B) \bar{y}_k] = 0 \right\}. \end{aligned} \quad (36b)$$

Using Lemma 3.2 and the fact that

$$\hat{t}_k = \beta_1 e_1 - B_{k+d} \hat{y}_k = \beta_1 e_1 - \overline{B}_{k+d} \bar{y}_k,$$

we obtain

$$\underline{\mu}_d(x_k, \theta) = \min\{\omega_k, \sigma_{\min}(N_{k,d})\}, \quad \bar{\mu}_d(x_k, \theta) = \min\{\omega_k, \sigma_{\min}(\overline{N}_{k,d})\}, \quad (37)$$

where

$$N_{k,d} \equiv [B_{k+d}, \omega_k(I - \hat{t}_k \hat{t}_k^\dagger)], \quad \overline{N}_{k,d} \equiv [\overline{B}_{k+d}, \omega_k(I - \hat{t}_k \hat{t}_k^\dagger)]. \quad (38)$$

Theorem 5.1. *Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and let $x_k = V_k y_k$ be the approximate solution to the least squares problem (1) computed by LSQR at the iteration step k . Let $\mu(x_k, \theta)$ be the minimal backward error defined in Lemma 3.2 associated to x_k for a given $\theta > 0$ and let $\underline{\mu}_d(x_k, \theta)$ and $\bar{\mu}_d(x_k, \theta)$ be defined by (37) for a given $d \geq 0$. Then*

$$\underline{\mu}_d(x_k, \theta) \leq \mu(x_k, \theta) \leq \bar{\mu}_d(x_k, \theta). \quad (39)$$

In addition,

$$0 = \underline{\mu}_0(x_k, \theta) \leq \cdots \leq \underline{\mu}_l(x_k, \theta) = \mu(x_k, \theta) = \bar{\mu}_l(x_k, \theta) \leq \cdots \leq \bar{\mu}_0(x_k, \theta) \quad (40)$$

for some $l \leq n - k$.

Proof. First we prove the lower bound in (39). If $\sigma_{\min}(N_k) > \omega_k$ then $\underline{\mu}_d(x_k, \theta) \leq \mu(x_k, \theta)$ holds since

$$\underline{\mu}_d(x_k, \theta) = \min\{\omega_k, \sigma_{\min}(N_{k,d})\} \leq \omega_k = \min\{\omega_k, \sigma_{\min}(N_k)\} = \mu(x_k, \theta).$$

On the other hand assume that $\sigma_{\min}(N_k) \leq \omega_k$, i.e., $\mu(x_k, \theta) = \sigma_{\min}(N_k)$. Let \tilde{U} and \tilde{V} be such that $U = [U_{k+d+1}, \tilde{U}]$ and $V = [V_{k+d}, \tilde{V}]$ are orthogonal matrices. Then using (6b) and (13),

$$U^T N_k \begin{bmatrix} V & 0 \\ 0 & U \end{bmatrix} = \left[\begin{array}{cc|cc} B_{k+d} & C & \omega_k(I - \hat{t}_k \hat{t}_k^\dagger) & 0 \\ 0 & D & 0 & \omega_k I \end{array} \right], \quad (41)$$

where $C = U_{k+d+1}^T A \tilde{V}$ and $D = \tilde{U}^T A \tilde{V}$. Owing to the interlacing property of singular values, see, e.g., [13, Corollary 3.1.3], the smallest singular value of N_k is not less than the smallest singular value of the matrix

$$\left[\begin{array}{cc|cc} B_{k+d} & \omega_k(I - \hat{t}_k \hat{t}_k^\dagger) & 0 & \\ 0 & 0 & \omega_k I & \end{array} \right] = \begin{bmatrix} N_{k,d} & 0 \\ 0 & \omega_k I \end{bmatrix}$$

obtained from (41) after deleting the columns corresponding to the matrices C and D . Hence

$$\mu(x_k, \theta) = \min\{\omega_k, \sigma_{\min}(N_k)\} = \sigma_{\min}(N_k) \geq \min\{\omega_k, \sigma_{\min}(N_{k,d})\} = \underline{\mu}_d(x_k, \theta)$$

due to the assumption $\sigma_{\min}(N_k) \leq \omega_k$. The lower bounds (40) on $\mu(x_k, \theta)$ follow from the interlacing property as well. Since y_k is the exact solution of the least squares problem (8), it follows that $\underline{\mu}_0(x_k, \theta) = 0$, while the existence of l such that $\underline{\mu}_l(x_k, \theta) = \mu(x_k, \theta)$ follows from the finite termination property of LSQR.

To prove the upper bound in (39) we have

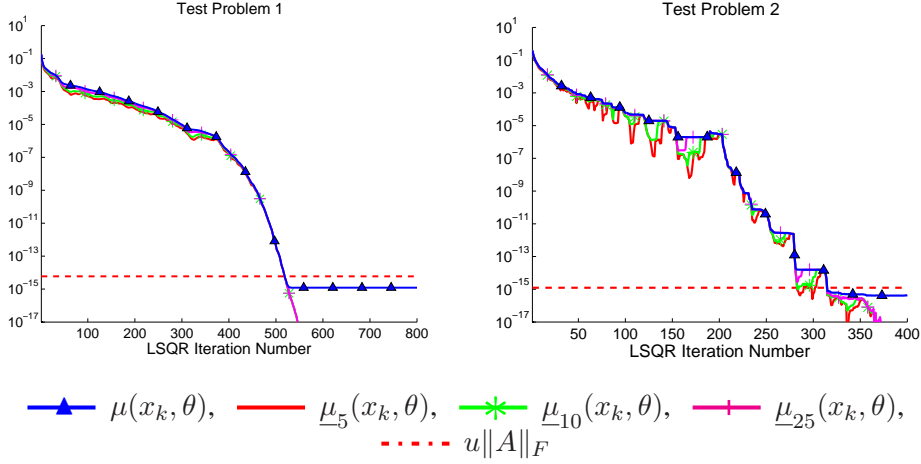
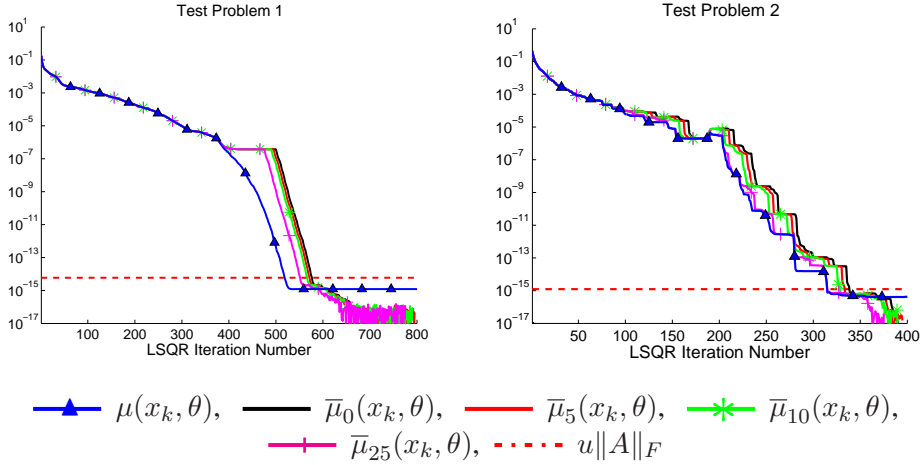
$$\begin{aligned} \sigma_{\min}(N_k) &= \min_{z; \|z\|_2=1} \left\| \begin{bmatrix} A^T \\ \omega_k(I - r_k r_k^\dagger) \end{bmatrix} z \right\|_2 \leq \min_{z; \|z\|_2=1} \left\| \begin{bmatrix} A^T \\ \omega_k(I - r_k r_k^\dagger) \end{bmatrix} U_{k+d+1} \tilde{z} \right\|_2 \\ &= \min_{\tilde{z}; \|\tilde{z}\|_2=1} \left\| \begin{bmatrix} A^T U_{k+d+1} \\ U_{k+d+1} \omega_k(I - \hat{t}_k \hat{t}_k^\dagger) \end{bmatrix} \tilde{z} \right\|_2 \\ &= \min_{\tilde{z}; \|\tilde{z}\|_2=1} \left\| \begin{bmatrix} V_{k+d+1} \bar{B}_{k+d}^T \\ U_{k+d+1} \omega_k(I - \hat{t}_k \hat{t}_k^\dagger) \end{bmatrix} \tilde{z} \right\|_2 \\ &= \min_{\tilde{z}; \|\tilde{z}\|_2=1} \left\| \begin{bmatrix} \bar{B}_{k+d}^T \\ \omega_k(I - \hat{t}_k \hat{t}_k^\dagger) \end{bmatrix} \tilde{z} \right\|_2 = \sigma_{\min}(\bar{N}_{k,d}). \end{aligned}$$

It hence follows that $\mu(x_k, \theta) \leq \bar{\mu}_d(x_k, \theta)$. The upper bounds (40) follow from the fact that with growing d we are enlarging the subspace $\text{range}(U_{k+d+1})$ over which we are minimizing to obtain the upper bounds. Since $\bar{N}_{k,d}$ differs from $N_{k,d}$ only by a zero column, we have the equality $\mu(x_k, \theta) = \bar{\mu}_l(x_k, \theta)$. \square

In order to evaluate the estimates $\underline{\mu}_d(x_k, \theta)$ and $\bar{\mu}_d(x_k, \theta)$ defined by (36) and (37), we need to be able to efficiently compute the smallest singular value of the matrices $N_{k,d}$ and $\bar{N}_{k,d}$ in (38). Let $\hat{Q}_k = \text{diag}(Q_k, I_d)$, where Q_k is the matrix of k reflections from (11). Using (13), the matrix $N_{k,d}$ is orthogonally similar to the matrix

$$M_{k,d} \equiv \hat{Q}_k N_{k,d} \begin{bmatrix} I_{k+d} & 0 \\ 0 & \hat{Q}_k^T \end{bmatrix} = [\hat{Q}_k B_{k+d}, \omega_k(I - e_{k+1} e_{k+1}^T)],$$

where e_{k+1} is the $(k+1)$ -st column of I_{k+d+1} here. The matrix $\hat{Q}_k B_{k+d}$ is the partially transformed

Figure 3: Estimating $\mu(x_k, \theta)$ using $\underline{\mu}_d(x_k, \theta)$ with various values of d .Figure 4: Estimating $\mu(x_k, \theta)$ using $\bar{\mu}_d(x_k, \theta)$ with various values of d .

5.2 An asymptotic estimate of $\mu(x_k, \theta)$

The literature (see [9], [10], [14], and [24]) indicate that the following quantity can be used as an estimate of $\mu(x_k, \infty)$:

$$\nu(x_k, \infty) \equiv \left\| \left(\|x_k\|_2^2 A^T A + \|r_k\|_2^2 I \right)^{-1/2} A^T r_k \right\|_2. \quad (43)$$

It is easy to show (see for example [24] or [14, §2]) that

$$\nu(x_k, \infty) = \left\| \begin{bmatrix} A \\ (\|r_k\|_2 / \|x_k\|_2) I \end{bmatrix} \begin{bmatrix} A \\ (\|r_k\|_2 / \|x_k\|_2) I \end{bmatrix}^\dagger \begin{bmatrix} r_k / \|x_k\|_2 \\ 0 \end{bmatrix} \right\|_2.$$

Note that $\nu(x_k, \infty)$ is a projection, much like the projection $\|P_A r_k\|_2 = \|A A^\dagger r_k\|_2$ discussed in Section 4.

Recall from Lemma 3.3 that we are interested in estimating $\mu(x_k, \hat{\theta})$ and not $\mu(x_k, \infty)$. Su [24, §2.7] extended the asymptotic estimate $\nu(x_k, \infty)$ to estimate $\mu(x_k, \theta)$ for any finite $\theta > 0$:

$$\nu(x_k, \theta) \equiv \left\| \begin{bmatrix} A \\ \omega_k I \end{bmatrix} \begin{bmatrix} A \\ \omega_k I \end{bmatrix}^\dagger \begin{bmatrix} \omega_k r_k / \|r_k\|_2 \\ 0 \end{bmatrix} \right\|_2, \quad (44)$$

where, as previously, ω_k is defined in Lemma 3.2. It is straightforward to verify that many relations between $\mu(x_k, \infty)$ and $\nu(x_k, \infty)$ also hold between $\mu(x_k, \theta)$ and $\nu(x_k, \theta)$. In particular, $\mu(x_k, \theta)$ and $\nu(x_k, \theta)$ are asymptotically equivalent:

$$\lim_{x_k \rightarrow \hat{x}} \frac{\mu(x_k, \theta)}{\nu(x_k, \theta)} = 1. \quad (45)$$

In Section 4 we discussed ways to estimate $\|P_{Ar_k}\|_2$ efficiently in LSQR, and we can proceed in a similar way to estimate the projection $\nu(x_k, \theta)$. Applying Lemma 4.1 to (44) gives

$$\underline{\nu}(x_k, \theta) \leq \nu(x_k, \theta) \leq \bar{\nu}(x_k, \theta), \quad (46)$$

where

$$\underline{\nu}(x_k, \theta) \equiv \frac{\omega_k}{\sqrt{\omega_k^2 + \sigma_{\max}^2(A)}} \frac{\|A^T r_k\|_2}{\|r_k\|_2}, \quad \bar{\nu}(x_k, \theta) \equiv \frac{\omega_k}{\sqrt{\omega_k^2 + \sigma_{\min}^2(A)}} \frac{\|A^T r_k\|_2}{\|r_k\|_2}.$$

It immediately follows from the above that

$$\frac{\bar{\nu}(x_k, \theta)}{\underline{\nu}(x_k, \theta)} = \sqrt{\frac{\omega_k^2 + \sigma_{\max}^2(A)}{\omega_k^2 + \sigma_{\min}^2(A)}}. \quad (47)$$

We can interpret (47) as follows. The bounds on $\nu(x_k, \theta)$ in (46) are tight when $\omega_k \gg \sigma_{\max}(A)$. They are also tight when $\omega_k \approx \sigma_{\max}(A)$, as well as when $\omega_k \ll \sigma_{\max}(A)$ provided $\kappa_2(A)$ is not too large. These new bounds on the asymptotic estimate only fail to be tight when $\omega_k \ll \sigma_{\max}(A)$ and $\kappa_2(A) \gg 1$.

The computation of the lower bound $\underline{\nu}(x_k, \theta)$ requires essentially no extra cost in LSQR, since reliable estimates of $\|r_k\|_2$, $\|A^T r_k\|_2$, and $\sigma_{\max}(A)$ are available at essentially no extra cost in LSQR. As in (26), the upper bound $\bar{\nu}(x_k, \theta)$ is much harder to compute, since it involves the smallest singular value of A .

We illustrate these bounds in Figure 5. As in Figures 3 and 4 we have set $\theta = \|A\|_F / \|b\|_2$. The asymptotic estimate $\nu(x_k, \theta)$ seems to be an excellent estimate of $\mu(x_k, \theta)$ not only asymptotically as $x_k \rightarrow \hat{x}$, but even in the first iterations when \hat{x} is very far from x_k . As with the bounds on $\|P_{Ar_k}\|_2$ in (26), the lower bound $\underline{\nu}(x_k, \theta)$ appears to be usually tighter than $\bar{\nu}(x_k, \theta)$, especially late in the iteration process. Both bounds oscillate in ill-conditioned problems. In well-conditioned problems, however, the lower bound $\underline{\nu}(x_k, \theta)$ seems to give good order-of-magnitude estimates of $\nu(x_k, \theta)$ and $\mu(x_k, \theta)$.

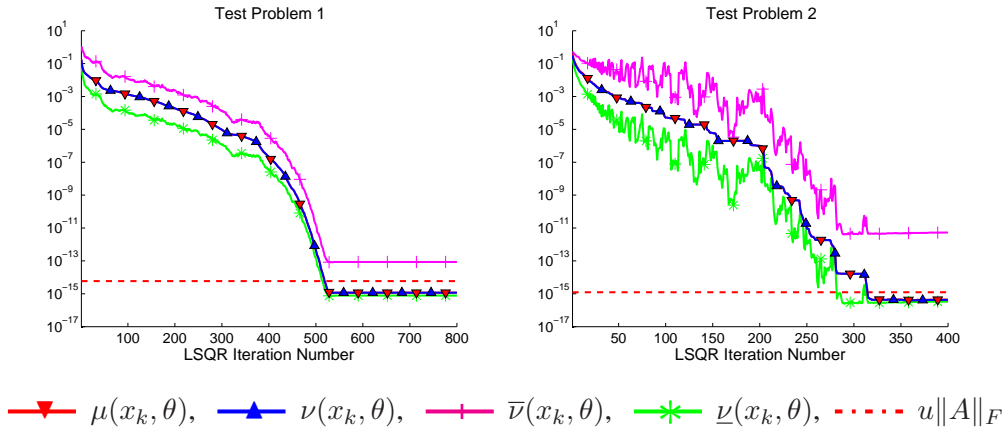


Figure 5: The bounds $\underline{\nu}(x_k, \theta)$ and $\bar{\nu}(x_k, \theta)$ on $\nu(x_k, \theta)$.

Using (33) and taking the 2-norm, we obtain

$$\underline{\mu}_d(x_k, \theta) = \frac{\omega_k}{\|\hat{t}_k\|_2} \left\| \begin{bmatrix} I_{k+d} & \\ & 0 \end{bmatrix} G_{k+d} [0_k, \phi_{k+1}, \dots, \phi_{k+d}, \bar{\phi}_{k+d+1}, 0_{k+d}]^T \right\|_2. \quad (50)$$

Thus the estimate $\underline{\mu}_d(x_k, \theta)$ can be computed by applying $2(k+d) - 1$ Givens rotations to the vector $[0_k, \phi_{k+1}, \dots, \phi_{k+d}, \bar{\phi}_{k+d+1}, 0_{k+d}]^T$ and computing the 2-norm of the vector formed by its first $k+d$ elements. The cost of this computation is $\mathcal{O}(k+d)$ flops and storage.

The estimate $\bar{\nu}_d(x_k, \theta)$ can be computed in a similar way. Instead of (49) we have

$$\begin{bmatrix} Q_{k+d} & \\ & I_{k+d+1} \end{bmatrix} \begin{bmatrix} \bar{B}_{k+d} \\ \omega_k I_{k+d+1} \end{bmatrix} = \left[\begin{array}{cccc|c} \rho_1 & \theta_2 & & & \\ & \rho_2 & \ddots & & \\ & & \ddots & \theta_{k+d} & \\ & & & \rho_{k+d} & \theta_{k+d+1} \\ 0 & 0 & \dots & 0 & \bar{\rho}_{k+d+1} \\ \omega_k & & & & \\ & \omega_k & & & \\ & & \ddots & & \\ & & & \omega_k & \\ & & & & \omega_k \end{array} \right], \quad (51)$$

which differs from C_{k+d} by only one extra column. Thus $\bar{\mu}_d(x_k, \theta)$ can be computed much in the same way as $\underline{\mu}_d(x_k, \theta)$, using only two extra Givens rotations. As with $\underline{\mu}_d(x_k, \theta)$ and $\bar{\mu}_d(x_k, \theta)$, the estimates $\underline{\nu}_d(x_k, \theta)$ and $\bar{\nu}_d(x_k, \theta)$ must be recomputed at every iteration, since they require the quantity ω_k that generally changes at each step.

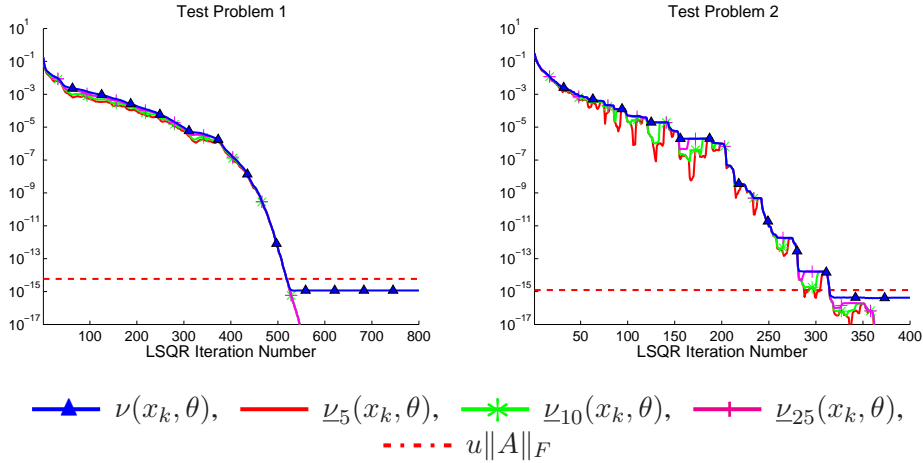


Figure 6: Estimating $\nu(x_k, \theta)$ using $\underline{\nu}_d(x_k, \theta)$ with various values of d .

The estimates $\underline{\nu}_d(x_k, \theta)$ and $\bar{\nu}_d(x_k, \theta)$ are compared with $\nu(x_k, \theta)$ in Figures 6 and 7 for various values of the parameter d and $\theta = \|A\|_F / \|b\|_2$. The behaviour of these estimates is very similar to that of $\underline{\mu}_d(x_k, \theta)$ and $\bar{\mu}_d(x_k, \theta)$ in Figures 3 and 4, not only asymptotically as $x_k \rightarrow \hat{x}$, but also very early in the iteration process.

6 Conclusion

In this paper we have proposed efficient methods to estimate the quantities $\|P_{AT}r_k\|_2$ in (19) and $\mu(x_k, \theta)$ in (22), which are both much too expensive to compute to be used directly in large sparse applications.

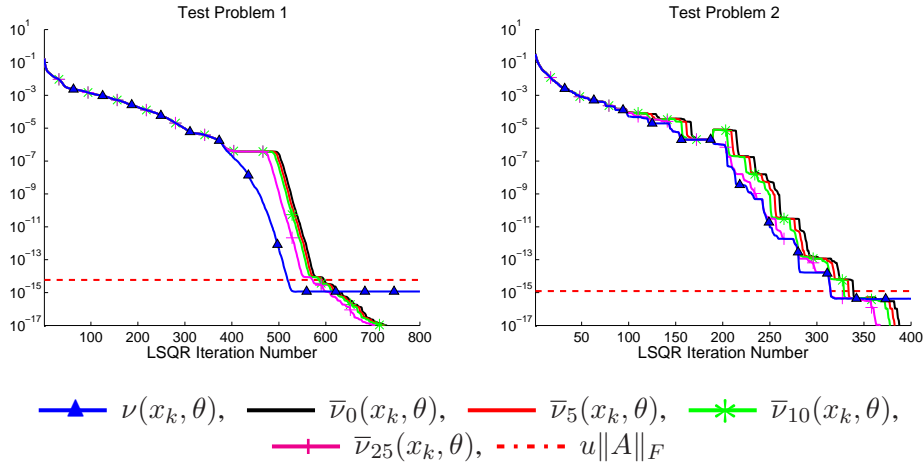


Figure 7: Estimating $\nu(x_k, \theta)$ using $\bar{\nu}_d(x_k, \theta)$ with various values of d .

Although we have focussed our discussion on the algorithm LSQR, these results could be applicable to other conjugate gradients type algorithms, such as CGLS.

In our experience, the bound (19) involving $\|P_{Ar_k}\|_2$ is often tight. It should be noted, however, that in some problems, in particular when the right-hand side vector is not well correlated with the data, the stopping criterion in (25) based on $\mu(x_k, \theta)$ can be triggered much earlier than that in (21) involving $\|P_{Ar_k}\|_2$, which is generally exact only asymptotically—compare (19) and (20) with (24). This justifies the use of the estimates of $\mu(x_k, \theta)$ in (25), rather than of $\|P_{Ar_k}\|_2$ in (21), as stopping criteria in some cases.

The quantity $\lambda_d(x_k)$ in (30) is a very practical lower bound on $\|P_{Ar_k}\|_2$. It can be computed at step $k + d$ of LSQR at essentially no extra cost.

The minimal backward error $\mu(x_k, \theta)$ can be estimated using the bounds $\underline{\mu}_d(x_k, \theta)$ and $\bar{\mu}_d(x_k, \theta)$ given in (36) and (37). One can also consider the asymptotic estimates $\underline{\nu}_d(x_k, \theta)$ and $\bar{\nu}_d(x_k, \theta)$ given in (48). All of these estimates can be computed at the step $k + d$ at a cost of $\mathcal{O}(k + d)$ flops and storage. Although this is slightly less efficient than the computation of $\lambda_d(x_k)$ (and not as trivial to implement) it is certainly feasible for practical large sparse problems. One might choose not to recompute these estimates at each iteration, but rather at the step when another stopping criterion, e.g.,

$$\lambda_d(x_k) \leq \alpha\|A\|_{2,F}\|x_k\|_2 + \beta\|b\|_2$$

is triggered.

All the estimates discussed above “lag behind” by d iterations, in other words, they can only be used to estimate $\|P_{Ar_k}\|_2$ or $\mu(x_k, \theta)$ at the iteration number $k + d$. To avoid such a delay, one can consider the upper bound $\bar{\mu}_0(x_k, \theta)$ on $\mu(x_k, \theta)$, as well as its asymptotic estimate $\bar{\nu}_0(x_k, \theta)$. One can also use Lemma 4.1 to estimate $\|P_{Ar_k}\|_2$ with its lower bound $\|A^T r_k\|_2/\|A\|_2$ and $\mu(x_k, \theta)$ with the lower bound $\underline{\nu}(x_k, \theta)$ on its asymptotic estimate; see (26) and (46). Both these estimates are available essentially free at step k of LSQR. Although they tend to oscillate in ill-conditioned problems, they are usually fairly tight and give good order of magnitude estimates in well-conditioned problems.

The estimates of $\|P_{Ar_k}\|_2$ and $\mu(x_k, \theta)$ discussed in this paper can be used in (21) and (25). As a result of this work, (21) and (25) can now be used as stopping criteria for the iterative solution of practical large sparse LS problems.

Acknowledgments

The authors are very grateful to Xiao-Wen Chang, Chris Paige, and Miro Rozložník. Their comments and suggestions have greatly improved the paper.

References

- [1] M. Arioli and S. Gratton. Least-squares problems, normal equations, and stopping criteria for the conjugate gradient method. Technical Report RAL-TR-2008-008, Rutherford Appleton Laboratory, 2008. [5](#), [8](#)
- [2] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia PA, 1996. [2](#), [3](#)
- [3] Å. Björck, T. Elfving, and Z. Strakoš. Stability of conjugate gradient and Lanczos methods for linear least squares problems. *SIAM J. Matrix Anal. Appl.*, 19(3):720–736, 1998. [2](#), [9](#)
- [4] R. F. Boisvert, R. Pozo, K. Remington, R. Barret, and J. J. Dongarra. The Matrix Market: A web resource for test matrix collections web resource for test matrix collections. In R. F. Boisvert, editor, *Quality of Numerical Software, Assessment and Enhancement*. Chapman & Hall, London, UK, 1997. [2](#)
- [5] X.-W. Chang, C. C. Paige, and D. Titley-Peloquin. Stopping criteria for the iterative solution of linear least squares problems. *SIAM J. Matrix Anal. Appl.*, 31(2):831–852, 2009. [2](#), [4](#), [6](#), [7](#), [9](#)
- [6] J. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, 1997. [14](#)
- [7] G. H. Golub and W. M. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *SIAM J. Numer. Anal. Ser. B*, 2(2):205–224, 1965. [3](#)
- [8] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore MD, third edition, 1996. [2](#)
- [9] J. F. Grcar. Optimal sensitivity analysis of linear least squares. Technical Report LBNL-52434, Lawrence Berkeley National Laboratory, 2003. [6](#), [10](#), [15](#)
- [10] M. Gu. Backward perturbation bounds for linear least squares problems. *SIAM J. Matrix Anal. Appl.*, 20(2):363–372, 1998. [6](#), [15](#)
- [11] P. C. Hasen. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. SIAM, Philadelphia PA, 1998. [17](#)
- [12] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Stand.*, 49(6):409–435, 1952. [2](#), [8](#)
- [13] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1991. [12](#)
- [14] R. Karlson and B. Waldén. Estimation of optimal backward perturbation bounds for the linear least squares problem. *BIT*, 37(4):862–869, 1997. [15](#)
- [15] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Stand.*, 45(4):255–282, 1950. [3](#)
- [16] C. C. Paige. Bidiagonalization of matrices and solution of linear equations. *SIAM J. Numer. Anal.*, 11(1):197–209, 1974. [4](#)
- [17] C. C. Paige and M. A. Saunders. Algorithm 583, LSQR: sparse linear equations and sparse least squares problems. *ACM Trans. Math. Software*, 8(2):195–209, 1982. [2](#), [3](#)
- [18] C. C. Paige and M. A. Saunders. LSQR: an algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Software*, 8(1):43–71, 1982. [2](#), [3](#), [4](#), [5](#)

-
- [19] J. L. Rigal and J. Gaches. On the compatibility of a given solution with the data of a linear system. *J. ACM*, 14(3):543–548, 1967. 5
- [20] G. W. Stewart. Research, development, and LINPACK. In *Mathematical Software III*, pages 1–14. Academic Press, New York, 1977. 5
- [21] G. W. Stewart and J.-g. Sun. *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic Press, San Diego CA, 1990. 5
- [22] Z. Strakoš and P. Tichý. On error estimation in the conjugate gradient method and why it works in finite precision computations. *Electron. Trans. Numer. Anal.*, 13:56–80, 2002. 8
- [23] Z. Strakoš and P. Tichý. Error estimation in preconditioned conjugate gradients. *BIT*, 45(4):789–817, 2005. 8
- [24] Z. Su. *Computational Methods for Least Squares Problems and Clinical Trials*. PhD thesis, Stanford University, 2005. 15
- [25] B. Waldén, R. Karlson, and J.-g. Sun. Optimal backward perturbation bounds for the linear least squares problem. *Numer. Linear Algebra Appl.*, 2(3):271–286, 1995. 6