Notes 3: Iterative Methods for Square and Rectangular Systems

1 Introduction

Some unsymmetric or rectangular systems that arise in optimization are:

- Simplex method: \( Bx = b \) and \( B^T y = c \), where \( B \) is a square basis matrix.
- First-order multiplier estimates: \( \min_y \| J^T y - g \| \), where \( J \) is the Jacobian for the current set of active constraints, and \( g \) is the current objective gradient.

Large linear programs usually require the solution of so many square systems that we could not consider using iterative methods to solve \( Bx = b \) and \( B^T y = c \) within the simplex method. However, special circumstances may arise requiring nonstandard approaches.

For various applications we consider iterative methods for the following problems:

- Square unsymmetric systems: \( Ax = b \).
- Under-determined consistent systems: \( \min \| x \|_2 \) subject to \( Ax = b \).
- Over-determined systems (least squares): \( \min \| Ax - b \|_2 \).
- Regularized systems: \( \min \| Ax - b \|_2^2 + \| \delta x \|_2^2 \) \( \equiv \min \| (A \delta I)x - (b_0) \|_2^2 \), where \( \delta \) is a scalar and \( A \) may have any shape or rank.

These four problems are equivalent to the symmetric system \( (\gamma I A A^T \delta I)x = (b_0) \) with the parameters \( (\gamma, \delta) = (0, 0), (0, -1), (1, 0), (\delta, -\delta) \) respectively. They are also equivalent to the definite or semidefinite systems \( (A^T A + \gamma^2 I)x = A^T b \) or \( (AA^T + \gamma^2 I)y = b \) with \( \gamma = 0 \) or \(-\delta \) respectively, and \( x = A^T y \). We may therefore expect the symmetric iterative solvers to apply. However, more effective numerical methods are obtained by working with \( A \) directly.

2 Bidiagonalization methods for unsymmetric systems

A square or rectangular matrix \( A \) can be reduced to upper bidiagonal form by multiplying alternately on the left and right by certain orthogonal matrices: \( U^T A V = B \). This is the starting point for dense singular value decompositions (SVDs) [7]. When \( A \) is sparse or a “black box” operator for forming matrix-vector products \( Av, A^T u \), the bidiagonalization can be performed iteratively. (Actually, we reduce \( b \ A \) to upper bidiagonal form, meaning \( A \) is reduced to lower bidiagonal form.)

2.1 The Golub-Kahan process

Bidiag\((A, b) \rightarrow (B_k, U_{k+1}, V_k) \) or \((L_k, U_k, V_k) \) denotes the following process. Given an \( m \times n \) matrix \( A \) and a starting vector \( b \), the Golub-Kahan process [5] generates vectors \( u_k, v_k \) and positive scalars \( \alpha_k, \beta_k \) \( (k = 1, 2, \ldots) \) according to these steps:

1. Set \( \beta_1 u_1 = b \) and \( \alpha_1 v_1 = A^T u_1 \). (Exit if \( \beta_1 = 0 \) or \( \alpha_1 = 0 \).)
2. For \( k = 1, 2, \ldots \), set

\[
\begin{align*}
\beta_{k+1} u_{k+1} &= A v_k - \alpha_k u_k, \\
\alpha_{k+1} v_{k+1} &= A^T u_{k+1} - \beta_{k+1} v_k.
\end{align*}
\]
After \( k \) steps, the situation is summarized by the equations

\[
AV_k = U_{k+1}B_k = U_kL_k + \beta_{k+1}u_{k+1}^T,
\]
\[
A^TU_{k+1} = V_kB_k^T + \alpha_{k+1}v_{k+1}^T = V_{k+1}L_{k+1}^T,
\]
where \( U_k = (u_1 \ u_2 \ldots \ u_k), V_k = (v_1 \ v_2 \ldots \ v_k), L_k \) is lower bidiagonal, and \( B_k \) is also lower bidiagonal with one extra row:

\[
L_k = \begin{pmatrix}
\alpha_1 & \alpha_2 & \ldots \\
\beta_2 & \alpha_2 & \ldots \\
& \ddots & \ddots \\
& & \beta_k & \alpha_k
\end{pmatrix}, \quad B_k = \begin{pmatrix}
\alpha_1 & \alpha_2 & \ldots \\
\beta_2 & \alpha_2 & \ldots \\
& \ddots & \ddots \\
& & \beta_k & \beta_{k+1}
\end{pmatrix} = \begin{pmatrix}
L_k \\
\beta_{k+1}e_k^T
\end{pmatrix}.
\]

With exact arithmetic the columns of \( U_k \) and \( V_k \) would be orthonormal for each \( k \) until \( \beta_{\ell+1} = 0 \) or \( \alpha_{\ell+1} = 0 \) for some \( k = \ell \). In practice, orthonormality is soon lost, but relations (1)–(2) hold to working precision. To retain orthogonality properties, one might guess that each \( u_{k+1} \) and \( v_{k+1} \) would need to be reorthogonalized with respect to \( U_k \) and \( V_k \) respectively, at the expense of storing all previous Golub-Kahan vectors. However, Simon and Zha [21] found that one-sided reorthogonalization is enough; that is, if \( u_{k+1} \) is reorthogonalized with respect to \( U_k \), then \( V_k \) remains essentially orthonormal, and vice versa. The same effect was observed by Fang and Saunders [3]. The Fortran 90 implementation of LSMR allows local or full reorthogonalization of \( V_k \) (the shortest vectors for over-determined systems).

A compromise is to use partial reorthogonalization, as in Larsen’s PROPACK package for computing some of the singular vectors of a sparse matrix or linear operator \( A \) [15].

### 2.2 Properties of the Golub-Kahan process

The vector \( u_k \) lies in the Krylov subspace \( K_k(AA^T, b) \equiv \text{span}\{b, AA^Tb, \ldots, (AA^T)^{k-1}b\} \), and \( v_k \) lies in the Krylov subspace \( K_k(A^TA, A^Tb) \equiv \text{span}\{A^Tb, (A^TA)A^Tb, \ldots, (A^TA)^{k-1}A^Tb\} \).

Some properties follow:

1. \( B_k \) has full column rank \( k \) for all \( k \leq \ell \).
2. If \( \beta_{\ell+1} = 0 \), we have \( AV_\ell = U_\ell L_\ell \) with \( L_\ell \) nonsingular and \( b \notin \text{range}(A) \).
3. If \( \beta_{\ell+1} > 0 \) but \( \alpha_{\ell+1} = 0 \), we have \( AV_\ell = U_{\ell+1}B_\ell \) with \( \text{rank}(B_\ell) = \ell \) but \( b \notin \text{range}(A) \).

### 2.3 Golub-Kahan with regularization

Let \( \delta \) be a given scalar (\( \geq 0 \) without loss of generality), and define

\[
\tilde{A} = \begin{pmatrix}
A \\
\delta I
\end{pmatrix}, \quad \tilde{b} = \begin{pmatrix}
b \\
0
\end{pmatrix}.
\]

During \( \text{Bidiag}(A, b) \), orthogonal matrices \( \tilde{Q}_k \) may be constructed from \( 2k-1 \) plane rotations to form the quantities

\[
\tilde{Q}_k \begin{pmatrix}
B_k \\
\delta I
\end{pmatrix} = \begin{pmatrix}
\tilde{B}_k \\
0
\end{pmatrix}, \quad \begin{pmatrix}
\tilde{U}_{k+1} & \tilde{Y}_k
\end{pmatrix} = \begin{pmatrix}
U_{k+1} & V_k
\end{pmatrix} \tilde{Q}_k^T,
\]

where \( \tilde{B}_k \) (like \( B_k \)) is lower bidiagonal with dimensions \((k+1) \times k\). The following result is obtained straightforwardly from (1)–(4).

**Result 2** If \( \text{Bidiag}(A, b) \to (B_k, U_{k+1}, V_k) \), then \( \text{Bidiag}(\tilde{A}, \tilde{b}) \to (\tilde{B}_k, \tilde{U}_{k+1}, \tilde{V}_k) \).

In short, the bidiagonalization of \( \tilde{A} = \begin{pmatrix} A \\ \delta I \end{pmatrix} \) may be obtained efficiently from the bidiagonalization of \( A \) itself. The mechanism is less trivial than in the symmetric case. It motivates the subproblems used next.
Table 4: Subproblems defining $y_k$ and $x_k = V_k y_k$ for three algorithms.

<table>
<thead>
<tr>
<th>Method</th>
<th>Subproblem</th>
<th>Factorization</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRAIG</td>
<td>$\delta = 0$</td>
<td>$L_k y_k = \beta_1 e_1$</td>
</tr>
<tr>
<td>LSQR</td>
<td>$\delta = 0$</td>
<td>$\min | B_k y_k - \beta_1 e_1 |$</td>
</tr>
<tr>
<td></td>
<td>$\delta &gt; 0$</td>
<td>$\min | (B_k \delta I) y_k - (\beta_1 e_1) |$</td>
</tr>
<tr>
<td>LSMR</td>
<td>$\delta \geq 0$</td>
<td>$\min | (R_k^T R_k) y_k - \bar{\beta}_1 e_1 |$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>$Q_k B_k = (R_k, 0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSQR</td>
<td>$Q_k (B_k \delta I) = (R_k, 0)$</td>
</tr>
<tr>
<td>LSMR</td>
<td>$Q_{k+1} (R_k^T) = (\bar{R}_k, 0)$</td>
</tr>
</tbody>
</table>

Table 5: Definition of $W_k$ and $z_k$ such that $x_k = V_k y_k = W_k z_k$.

<table>
<thead>
<tr>
<th>$W_k$</th>
<th>$z_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRAIG</td>
<td>$V_k$</td>
</tr>
<tr>
<td>LSQR</td>
<td>$V_k R_k^{-1}$</td>
</tr>
<tr>
<td>LSMR</td>
<td>$V_k R_k^{-1} \bar{R}_k^{-1}$</td>
</tr>
</tbody>
</table>

Table 4 shows how the factorizations are used to solve them. Table 5 shows how the factorizations needed to solve them. Table 4 shows the subproblems $LSMR$ method for solving unsymmetric $Ax$ method, where $A$ is the last diagonal element of $A$. Craig’s method (as derived by Paige [12]) is the simplest method for solving unsymmetric $Ax = b$, but not least-squares problems. $LSQR$ [13, 14] and $LSMR$ [3] apply to both consistent and inconsistent systems.

Since $\|U_{k+1}\| = O(1)$, we would like $t_{k+1}$ to be small (when $\delta = 0$). Craig’s method makes $t_{k+1} = 0$ everywhere except its last element, giving $r_k = -\eta_k \beta_{k+1} u_{k+1}$ (where $\eta_k$ is the last element of $y_k$), while $LSQR$ is more balanced in minimizing $\|t_{k+1}\|$ (which has the effect of minimizing $\|r_k\|$).

From (1) we have

$$ r_k = b - A x_k = \beta_1 u_1 - A V_k y_k = U_{k+1}(\beta_1 e_1 - B_k y_k) = U_{k+1} t_{k+1}, $$

where $t_{k+1} = \beta_1 e_1 - B_k y_k$.  \hspace{1cm} (5)

Since $\|U_{k+1}\| = O(1)$, we would like $t_{k+1}$ to be small (when $\delta = 0$). Craig’s method makes $t_{k+1} = 0$ everywhere except its last element, giving $r_k = -\eta_k \beta_{k+1} u_{k+1}$ (where $\eta_k$ is the last element of $y_k$), while $LSQR$ is more balanced in minimizing $\|t_{k+1}\|$ (which has the effect of minimizing $\|r_k\|$).

From (2) and (5)–(6) we have

$$ A^T r_k = A^T U_{k+1} t_{k+1} = V_{k+1} L_{k+1}^T t_{k+1} $$

$$ = V_{k+1} \left[ \left( \alpha_1 \beta_1 \right) - \left( B_k^T B_k \right) \right] y_k. $$

With $B_k^T B_k = R_k^T R_k$ and $\bar{\beta}_k = \alpha_k \beta_k$, we see that the $LSMR$ subproblem minimizes $\|A^T r_k\|$.

The $LSMR$ factorization needs $\varphi_k = \bar{\beta}_{k+1}/\rho_k$, where $\rho_k$ is the last diagonal of $R_k$.

**Result 3** CRAIG is equivalent to CG on $AA^T y = b$, where $x = A^T y$.

LSQR is equivalent to CG on the normal equation $(A^T A + \delta^2 I) x = A^T b$.

LSMR is equivalent to MINRES on the normal equation.
The CRAIG iteration  Given the Golub-Kahan process, CRAIG needs no further factorizations. Forward substitution on $L_k y_k = \beta_1 e_1$ gives $\eta_1 = \beta_1 / \alpha_1$, $\eta_2 = -\beta_2 \eta_1 / \alpha_2$, ..., and we have theoretically orthogonal updates to $x_0 \equiv 0$:

$$\eta_k = -\beta_k \eta_{k-1} / \alpha_k, \quad x_k = V_k y_k = x_{k-1} + \eta_k v_k.$$  

The LSQR iteration  The subproblem that allows LSQR to incorporate regularization was first proposed by Björck [1]. Result 2 helps confirm that the resulting method is equivalent to applying LSQR to $A$ and $b$. (Working backwards, the proof of Result 2 reveals the need for the orthogonal factorization (4), which in turn suggests the subproblem.) The QR factorizations for LSQR can be computed at negligible cost using one plane rotation for each $k$ when $\delta = 0$, or two rotations when $\delta > 0$, giving upper-bidiagonal factors $R_k$. The matrix $Q_k$, a product of the rotations, does not need to be saved. The columns of $W_k$ are obtained by forward substitution: $R_k^T W_k = V_k^T$. Then with $x_0 \equiv 0$,

$$x_k = V_k y_k = W_k R_k y_k = W_k z_k = x_{k-1} + \zeta_k w_k.$$  

The LSMR iteration  For LSMR we define $t_k = R_k y_k$ and solve $R_k^T q_k = \bar{\beta}_{k+1} e_k$ to get $q_k = (\bar{\beta}_{k+1} / \rho_k) e_k = \varphi_k e_k$ with $\rho_k = (R_k)_{kk}$ and $\varphi_k \equiv \bar{\beta}_{k+1} / \rho_k$. Then we perform a second QR factorization

$$\bar{Q}_{k+1} \begin{pmatrix} R_k^T & \bar{\beta}_{k+1} e_k \\ \varphi_k e_k^T & 0 \end{pmatrix} = \begin{pmatrix} \bar{R}_k & z_k \\ 0 & \bar{\zeta}_{k+1} \end{pmatrix}, \quad \bar{R}_k = \begin{pmatrix} \bar{\rho}_1 & \bar{\theta}_2 & \cdots \\ \bar{\theta}_2 & \ddots & \cdots \\ \bar{\theta}_k & \bar{\rho}_k \end{pmatrix}. \quad (8)$$

Combining with (7) gives

$$\min_{y_k} \| A^T r_k \| = \min_{y_k} \| \bar{\beta}_{k+1} e_1 - (R_k^T R_k) y_k \| = \min_{t_k} \| \bar{\beta}_{k+1} e_1 - (R_k^T \varphi_k e_k^T) \| t_k \|
= \min_{t_k} \| (z_k / \bar{\zeta}_{k+1}) - (\bar{R}_k / 0) t_k \|. \quad (9)$$

The subproblem is solved by choosing $t_k$ from $\bar{R}_k t_k = z_k$. Let $W_k$ and $\bar{W}_k$ be computed by forward substitution from $R_k^T W_k^T = V_k^T$ and $R_k^T \bar{W}_k^T = W_k^T$. Then from $x_k = V_k y_k$, $R_k y_k = t_k$, and $\bar{R}_k t_k = z_k$, we have $x_0 \equiv 0$ and

$$x_k = W_k R_k y_k = W_k t_k = \bar{W}_k \bar{R}_k t_k = \bar{W}_k z_k = x_{k-1} + \zeta_k \bar{w}_k.$$  

The following table compares the costs.

<table>
<thead>
<tr>
<th></th>
<th>Storage</th>
<th>Work per iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRAIG, $\delta = 0$</td>
<td>$m + 2n$</td>
<td>$3m + 4n$</td>
</tr>
<tr>
<td>LSQR, any $\delta$</td>
<td>$m + 3n$</td>
<td>$3m + 5n$</td>
</tr>
<tr>
<td>LSMR, any $\delta$</td>
<td>$m + 4n$</td>
<td>$3m + 6n$</td>
</tr>
</tbody>
</table>

2.5 Preferences  

Let $r_k = b - Ax_k$ be the residual and $d_k = x - x_k$ be the corresponding error. The properties of CRAIG are similar to those of SYMMLQ. The CRAIG point solves both of the problems

$$\begin{align*}
\min_{t_k} \| d_k \| \quad & \text{such that} \quad x_k = A^T U_k t_k, \\
\min_{y_k} \| x_k \| \quad & \text{such that} \quad x_k = V_k y_k, \quad U_k^T r_k = 0
\end{align*}$$
so that \(\|d_k\|\) decreases, \(\|x_k\|\) increases, and the system must be consistent. LSQR chooses \(y_k\) to solve the problem

\[
\min_{y_k} \|r_k\| \quad \text{such that} \quad x_k = V_k y_k,
\]

so that \(\|r_k\|\) decreases and the system may be inconsistent (\(\|r_k\| \neq 0\)). LSMR chooses \(y_k\) to solve

\[
\min_{y_k} \|A^T r_k\| \quad \text{such that} \quad x_k = V_k y_k,
\]

so that \(\|A^T r_k\|\) decreases, and again the system may be inconsistent.

Further discussion is given in \([19, 3]\). While CRAIG is slightly more economical, a conclusion in \([19]\) is that LSQR is suitable for all cases in the sense that it is reliable on square, over-determined, and under-determined systems, with or without regularization. We expect the same for LSMR. For consistent systems, LSQR is probably best because it minimizes \(\|r_k\|\). For least-squares problems, LSMR seems preferable if iterations are terminated early, because it is \(\|A^T r_k\|\) that must decrease to zero. A fortunate surprise is that LSMR’s \(\|r_k\|\) also decreases, and is never far behind \(\|r_k\|\) for LSQR.

Note that the solutions \(x_k = V_k y_k\) lie in the Krylov subspace \(K_k(A^T A, A^T b)\). The convergence of these bidiagonalization-based methods depends on the eigenvalues of \(A^T A\) (the squares of the singular values of \(A\)).

### 2.6 Estimation of norms

At iteration \(k\) of LSQR, estimates of \(\|r_k\|\), \(\|A^T r_k\|\), \(\|x_k\|\), \(\|A\|\), and \(\text{cond}(A)\) can be obtained at minimal cost. All five items are used in LSQR’s stopping rules. To estimate norms, it is often necessary to assume that the columns of \(U_k\) and \(V_k\) are orthonormal \((U_k^T U_k = I, V_k^T V_k = I)\). Although this is rarely true in practice, the resulting estimates have proved to be remarkably reliable.

For simplicity we assume \(\delta = 0\). As shown in \([13]\), the following relations can be derived from (1), (2), (5), (6), and the QR factorization of \(B_k\) in Table 4:

\[
\begin{align*}
   r_k &= \tilde{c}_{k+1} U_{k+1} Q_k^T e_{k+1} \\
   \|r_k\| &= \tilde{c}_{k+1} = \beta_1 s_1 s_2 \ldots s_k \\
   A^T r_k &= -(\tilde{c}_{k+1} \alpha_{k+1} c_k) v_{k+1} \\
   \|A^T r_k\| &= \tilde{c}_{k+1} \alpha_{k+1} |c_k|.
\end{align*}
\]

With orthogonality assumptions we also have \(V_k^T A^T A V_k = B_k^T B_k = R_k^T R_k\), and so from the Courant-Fischer minimax theorem, the eigenvalues of \(B_k^T B_k\) are bounded above and below by the largest and smallest nonzero eigenvalues of \(A^T A\). The same can be said of the singular values of \(B_k\) compared to those of \(A\). It follows for the 2- and F-norms that \(\|B_k\| \leq \|A\|\) and \(\|R_k^{-1}\| = \|B_k^{-1}\| \leq \|A^+\|\). With \(W_k = V_k R_k^{-1}\) we now have

\[
1 \leq \|B_k\| \|W_k\| \leq \|A\| \|A^+\| = \text{cond}(A)
\]

for the 2- and F-norms. Hence we use the monotonically increasing estimates

\[
\|A\|_F \approx \|B_k\|_F, \quad \text{cond}(A) \approx \|B_k\|_F \|W_k\|_F,
\]

where \(\|B_k\|_F^2 = \|B_k - 1\|_F^2 + \sigma_k^2 + \beta_{k+1}^2\) and \(\|W_k\|_F^2 = \|W_k - 1\|_F^2 + \|w_k\|^2\) can be updated easily.

To estimate \(\|x_k\|\), we make use of the so-called QLP factorization of \(B_k\) (see Stewart \([25]\)). The QR factorization of \(B_k\) can be followed by an orthogonal transformation from the right that reduces \(R_k\) to lower-bidiagonal form: \(R_k P_k = \hat{L}_k\). Recall that the least-squares subproblem for defining \(y_k\) is solved by \(R_k y_k = z_k\). If we define \(y_k = \hat{P}_k p_k\) for some vector
p_k$, we have $R_ky_k = R_kP_kp_k = \tilde{L}_k p_k = z_k$, where $p_k$ is obtained by forward substitution and hence differs from $p_{k-1}$ in just its last element $\pi_k$. We now have $\|x_k\|^2 = \|V_ky_k\|^2 \approx \|p_k\|^2$ and hence

$$\|x_k\|^2 = \|p_{k-1}\|^2 + \pi_k^2 = \|x_{k-1}\|^2 + \pi_k^2.$$  

This construction shows that $\|x_k\|$ increases monotonically for LSQR. The same approach applies to MINRES, and has been implemented in MINRES-QLP [2].

For LSMR, estimates of $\|r_k\|$, $\|A^T r\|$, $\|x_k\|$, $\|A\|$, and cond$(A)$ can also be obtained at negligible cost (although it is simpler to compute $\|x_k\|$ directly).

### 2.7 Stopping rules

We formulate rules for terminating LSQR and LSMR in terms of three dimensionless quantities $\alpha$, $\beta$, $\gamma$ specified by a user. (They are called atol, btol, conlim in the software.) The first two rules apply to consistent and inconsistent systems respectively. The third rule applies to both.

**S1** Stop if $\|r_k\| \leq \psi_k \equiv \alpha \|A\| \|x_k\| + \beta \|b\|$.

**S2** Stop if $\frac{\|A^T r_k\|}{\|A\| \|r_k\|} \leq \alpha$.

**S3** Stop if cond$(A) \geq \gamma$.

Rules S1 and S2 are based on allowable perturbations in the data. The user may therefore set $\alpha$ and $\beta$ according to the (known or estimated) accuracy of the data. Rule S3 represents an attempt to regularize ill-conditioned systems.

To justify S1, note that $x_k$ is the exact solution of the perturbed system

$$(A + E_k)x_k = b + f_k,$$

$$E_k = \frac{\alpha \|A\|}{\psi_k \|x_k\|} r_k x_k^T, \quad \|E_k\| = \frac{\alpha \|A\| \|r_k\|}{\psi_k},$$

$$f_k = -\frac{\beta \|b\|}{\psi_k} r_k, \quad \|f_k\| = \frac{\beta \|r_k\|}{\psi_k}$$

for any nonnegative $\alpha$ and $\beta$. If S1 is satisfied ($\|r_k\| \leq \psi_k$) we see that $x_k$ is an “acceptable solution” in the sense that $\|E_k\|/\|A\| \leq \alpha$ and $\|f_k\|/\|b\| \leq \beta$. Titley-Peloquin [26] shows that these $E_k$ and $f_k$ are the smallest perturbations to $A$ and $b$ that make $x_k$ acceptable.

To justify S2, Stewart [24] noted that $x_k$ and $\tilde{r}_k \equiv b - (A + E_k)x_k = r_k - E_k x_k$ are the exact solution and residual for the perturbed least-squares problem

$$\min_x \| (A + E_k)x - b \|, \quad E_k \equiv -\frac{r_k t_k^T A}{\|r_k\|^2}, \quad \|E_k\| = \|A^T r_k\|/\|r_k\|.$$  

Hence the perturbation to $A$ is sufficiently small ($x_k$ is an acceptable solution) if S2 is satisfied. This perturbation $E_k$ is not the smallest possible, but it is cheaply computable. We do know analytically the smallest perturbations to $A$ and $b$ that make $x_k$ an exact least-squares solution (Higham [9, pp. 392–393]), but normally they are much too expensive to evaluate.

Stopping rule S3 is based on the following arguments. Suppose that $A = USV^T$ (the SVD) with $S = \text{diag}(\sigma_j)$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. As $k$ increases, our estimate cond$(A) \approx \|B_k\| \|\psi\| |W_k|_r$ tends to level off near the values of the ordered sequence $\sigma_1/\sigma_1$, $\sigma_1/\sigma_2$, $\sigma_1/\sigma_3$, \ldots, with varying numbers of iterations near each level. This suggests rule S3 as a means of regularizing ill-conditioned problems, as in the discretization of ill-posed problems (e.g., [8]). For example, if the singular values of $A$ were known to be of order 1, 0.9, 10$^{-3}$, 10$^{-6}$, 10$^{-7}$, the effect of the two smallest singular values could be suppressed by setting $\gamma = 10^4$. 
3 Arnoldi-based methods for square systems

For square unsymmetric systems $Ax = b$, it may seem desirable to develop a method in which the approximate solutions $x_k = V_k y_k$ lie in the Krylov subspace $K_k(A, b)$ rather than $K_k(A^T A, A^T b)$ for LSQR and LSMR. Indeed this is possible, but at the cost of steadily increasing work and storage.

3.1 The Arnoldi process

Given a general $n \times n$ matrix $A$ and a starting vector $b$, the Arnoldi process generates vectors $v_k$ and scalars $\beta_k$ and $h_{ik}$ as follows:

1. Set $\beta_1 v_1 = b$. (Exit if $\beta_1 = 0$.)
2. For $k = 1, 2, \ldots, n$
   
   Compute $w = Av_k$
   
   For $i = 1, 2, \ldots, k$, set
   
   $h_{ik} = v_i^T w$
   
   $w = w - h_{ik} v_i$.
   
   Set $\beta_{k+1} v_{k+1} = w$. (If $\beta_{k+1} = 0$, define $\ell = k$ and exit.)

This is called the Modified Gram-Schmidt version of Arnoldi, and leads to the MGS-GMRES algorithm for solving $Ax = b$. As for the symmetric Lanczos process, each matrix $V_k = (v_1 \ v_2 \ \ldots \ v_k)$ would have orthonormal columns with exact arithmetic. After $k$ steps, the situation may be summarized as

$$AV_k = V_{k+1} H_k,$$

where $H_k$ is now a $(k+1) \times k$ Hessenberg matrix

$$H_k = \begin{pmatrix}
  h_{11} & h_{12} & \cdots & h_{1k} \\
  \beta_2 & h_{22} & \cdots & h_{2k} \\
  \beta_3 & \cdots & \cdots & h_{3k} \\
  \vdots & \vdots & & \vdots \\
  \beta_k & \cdots & \cdots & h_{kk} \\
  \beta_{k+1} & \cdots & \cdots & \cdots & h_{kk} \\
\end{pmatrix}.$$

The process stops with $k = \ell$ and $\beta_{\ell+1} = 0$ for some $\ell \leq n$, and then $AV_\ell = V_\ell T_\ell$, where $T_\ell$ is upper Hessenberg and square.

3.2 GMRES

If we define $x_k = V_k y_k$ for some $y_k$, (10) gives the residual

$$r_k = b - Ax_k = \beta_1 v_1 - AV_k y_k = V_{k+1} (\beta_1 e_1 - H_k y_k),$$

which suggests the least-squares subproblem $\min \| H_k y_k - \beta_1 e_1 \|$ for defining $y_k$ (as in MINRES and LSQR). This is the basis for the generalized minimum residual method GMRES [18]. For each $k$, a single plane rotation maintains the QR factorization

$$Q_k H_k = \begin{pmatrix} R_k & 0 \\ 0 & \end{pmatrix}, \quad Q_k (\beta_1 e_1) = \begin{pmatrix} z_k \\ \zeta_{k+1} \end{pmatrix},$$

from which the least-squares solution can be obtained when necessary by back-substitution: $R_k y_k = z_k$. Before that we have

$$r_k = V_{k+1} Q_k^T \left( \begin{pmatrix} z_k \\ \zeta_{k+1} \end{pmatrix} - \begin{pmatrix} R_k \\ 0 \end{pmatrix} y_k \right) = \zeta_{k+1} V_{k+1} Q_k^T e_{k+1},$$
so the estimate \( \|r_k\| \approx |y_{k+1}| \) is available without \( y_k \) or \( x_k \) (and as for MINRES and LSQR it is a reliable estimate in practice, even if the columns of \( V_k \) are not reorthogonalized).

Note that the Arnoldi process requires all columns of \( V_k \) to be retained. Whenever \( \|r_k\| \) is judged suitably small (or when \( k \) reaches a pre-defined limit \( m \)), the simplest form of GMRES terminates by solving \( R_k y_k = z_k \) and forming \( x_k = V_k y_k \) directly. There is no need to form \( x_1, x_2, \ldots, x_{k-1} \).

### 3.3 Restarted GMRES

For large systems, the iteration limit \( m \) is necessarily quite small (perhaps only 10 or 20). The simplest approach is to build a loop around GMRES analogous to iterative refinement for linear systems. The resulting algorithm is called GMRES(\( m \)):

1. Given \( x_0 \), form \( r = b - Ax_0 \).
2. If \( \|r\| \) is sufficiently small, exit with \( x = x_0 \).
3. Run the Arnoldi process for \( m \) iterations, starting with \( \beta_1 v_1 = r \).
4. Solve the least-squares problem \( \min \|H_m y_m - \beta_1 e_1\| \) and form \( \Delta x = V_m y_m \).
5. Set \( x_0 \leftarrow x_0 + \Delta x \) and return to step 1.

A difficulty with GMRES(\( m \)) is the need to choose \( m \).

If \( A \) is positive definite (the symmetric part \( (A + A^T)/2 \) is positive definite), it can be proved that GMRES(\( m \)) converges for any \( m \geq 1 \) [17].

If \( A \) is diagonalizable—i.e., it has a full set of eigenvalues and can be written as \( A = X D X^{-1} \) with \( D \) diagonal (but perhaps complex)—the rate of convergence can be bounded in terms of the condition number of \( X \). This is most useful if \( A \) is normal or nearly normal \( (A^T A \approx A A^T) \), in which case \( \text{cond}(X) \approx 1 \).

For more general matrices \( A \), GMRES is in danger of stagnation, in the sense that progress may become negligible for any practical value of \( m \). (This is in contrast to CRAIG, LSQR, and LSMR, which may converge slowly at times but will always get there if allowed to iterate long enough, and they do so using a constant amount of storage. The same is true of the symmetric methods CG, SYMMLQ, MINRES.) Nevertheless, GMRES is a remarkably popular method for solving unsymmetric square \( A x = b \). One advantage is that GMRES uses products \( A v \) but no transpose products \( A^T u \) (which may not be available). A further saving grace is the availability of effective preconditioners for many practical cases.

### 4 Preconditioning

It is a cliche to say that iterative methods need good preconditioners, but this is the main hope for minimizing the total iterations required, and for keeping \( m \) low for GMRES(\( m \)).

For square systems \( A x = b \), there is a choice of left, right, or split preconditioning. We need matrices \( C_1, C_2 \) such that \( C_1 C_2 \approx A \) and systems \( C_1 z = v \) and possibly \( C_2^T z = u \) can be solved efficiently \((i = 1, 2)\). The iterative method is then applied to \( C_1^{-1} A C_2^{-1} y = C_1^{-1} b_x \) and the solution is recovered by solving \( C_2 x = y \). The choice of \( C_i \) inevitably depends on each application.

In the absence of other knowledge, the very least we should do is apply row and column scaling. The general aim would be to ensure that all rows and all columns have essentially the same norm. Some efficient scaling algorithms are described in [10] for symmetric, square, and rectangular systems.

More generally, incomplete Cholesky or LU factorizations of \( A \) are commonly used, in which the “incomplete” factors are more sparse than the exact factors.

For least-squares problems, we must use \( C_1 = I \) to avoid altering the problem. Again, the very least we should do to help LSQR and LSMR is apply column scaling to make the 2-norms of all columns of \( A C_2^{-1} \) equal. Thus, \( C_2 \) should be a diagonal matrix whose \( j \)-th diagonal is \( \|a_j\|_2 \), where \( a_j \) is the corresponding column of \( A \). (This is diagonal preconditioning.)
More generally, \( C_2 \) should approximate the \( R \) part of a QR factorization of \( A \) (because the exact \( R \) would give convergence in one iteration—the perfect preconditioner). It is more likely that we could compute sparse rectangular LU factors \( P_1 A P_2 = (L_1 \ 0) \ (U_1 \ 0) \) with the \( L \) factor well-conditioned, and then \( C_2 = U P_2^T \) may be effective.

5 Other methods for unsymmetric systems

LSQR has been the primary iterative solver for over-determined (least-squares) problems, but LSMR will probably begin to take over. Both solvers should be kept in mind for square and under-determined systems.

Other important methods for square unsymmetric systems are USYMLQ and USYMQR \cite{20}, CGS \cite{22}, QMR \cite{4}, BiCGSTAB \cite{27}, and IDR(s) \cite{23}. Some intriguing test cases are given by Nachtigal et al. \cite{11} to show that LSQR, GMRES, and CGS are fundamentally different methods that can outperform each other (in iteration counts) by factors as large as \( \sqrt{n} \) or \( n \).

USYMLQ and USYMQR \cite{20} are based on an orthogonal tridiagonalization that becomes the symmetric Lanczos process when the matrix involved is symmetric. Two starting vectors \( b \) and \( c \) and products \( A v_k \) and \( A^T u_k \) are used to generate two orthonormal sets of vectors \( V_k = (v_1 \ldots v_k) \) and \( U_k = (u_1 \ldots u_k) \) for which

\[
\begin{align*}
(b \ AV_k) &= U_{k+1} (\beta e_1^T T_{k+1,k}^T), \\
(c \ A^T U_k) &= V_{k+1} (\gamma e_1^T T_{k+1,k}^T),
\end{align*}
\]

where \( T_{p,q} \in \mathbb{R}^{p \times q} \) is tridiagonal. Two systems \( Ax = b \) and \( A^T y = c \) can then be solved simultaneously. Reichel and Ye \cite{16} rediscovered the tridiagonalization and recognized that it applies equally well to rectangular systems. With square and rectangular systems in mind they named their solver GLSQR, even though it does not reduce to LSQR in any particular case. They showed that careful choice of \( c \) (namely \( c \approx x \)) can give good approximate solutions to \( Ax \approx b \) in fewer iterations than LSQR.

Soon after, Golub, Stoll, and Wathen \cite{6} used the orthogonal tridiagonalization to estimate the “scattering amplitude” \( c^T x = b^T y \) without computing \( x \) and \( y \).

5.1 Popularity contest

Google Scholar searches for the following sets of keywords

\[
\begin{array}{cccc}
\text{cg} & \text{hestenes stiefel} & \text{gmres} & \text{saad schultz} \\
\text{symmlq} & \text{paige saunders} & \text{qmr} & \text{freund nachtigal} \\
\text{minres paige saunders} & \text{bicgstab “van der Vorst”} & \text{lsqr paige saunders} & \text{Idr sonneveld “van Gijzen”}
\end{array}
\]

have given the following results in recent years (evidently exaggerated for MINRES).

<table>
<thead>
<tr>
<th>April</th>
<th>CG</th>
<th>SYMMLQ</th>
<th>MINRES</th>
<th>GMRES</th>
<th>QMR</th>
<th>BiCGSTAB</th>
<th>IDR(s)</th>
<th>LSQR</th>
<th>LSMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>1990</td>
<td>310</td>
<td>16100</td>
<td>5310</td>
<td>1080</td>
<td>1720</td>
<td>88</td>
<td>1780</td>
<td>6</td>
</tr>
<tr>
<td>2012</td>
<td>2200</td>
<td>324</td>
<td>16300</td>
<td>5480</td>
<td>1100</td>
<td>1870</td>
<td>117</td>
<td>1920</td>
<td>18</td>
</tr>
<tr>
<td>2013</td>
<td>2620</td>
<td>372</td>
<td>16700</td>
<td>6390</td>
<td>1220</td>
<td>2080</td>
<td>161</td>
<td>2390</td>
<td>38</td>
</tr>
<tr>
<td>2014</td>
<td>3070</td>
<td>387</td>
<td>2080</td>
<td>6860</td>
<td>1270</td>
<td>2270</td>
<td>168</td>
<td>2550</td>
<td>758</td>
</tr>
</tbody>
</table>

References


