Les Cahiers du GERAD ISSN: 0711–2440

MinAres : An iterative solver for symmetric linear systems

A. Montoison, D. Orban, M. A. Saunders

G–2023–40

August 2023

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2023
– Bibliothèque et Archives Canada, 2023

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2023
– Library and Archives Canada, 2023

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

MinAres : An iterative solver for symmetric linear systems

Alexis Montoison [a]
Dominique Orban [a]
Michael A. Saunders [b]

[a] *GERAD and Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal (Qc), Canada, H3T 1J4*

[b] *Systems Optimization Laboratory, Department of Management Science and Engineering, Stanford University, Stanford, CA, USA*

alexis.montoison@polymtl.ca
dominique.orban@gerad.ca
saunders@stanford.edu

August 2023
Les Cahiers du GERAD
G–2023–40

Abstract : We introduce an iterative solver named MinAres for symmetric linear systems $Ax \approx b$, where A is possibly singular. MinAres is based on the symmetric Lanczos process, like Minres and Minres-qlp, but it minimizes $\|Ar_k\|$ in each Krylov subspace rather than $\|r_k\|$, where r_k is the current residual vector. When A is symmetric, MinAres minimizes the same quantity $\|Ar_k\|$ as Lsmr, but in more relevant Krylov subspaces, and it requires only one matrix-vector product Av per iteration, whereas Lsmr would need two. Our numerical experiments with Minres-qlp and Lsmr show that MinAres is a pertinent alternative on consistent symmetric systems and the most suitable Krylov method for inconsistent symmetric systems. We derive properties of MinAres from an equivalent solver named cAr that is to MinAres as Cr is to Minres, is not based on the Lanczos process, and minimizes $\|Ar_k\|$ in the same Krylov subspace as MinAres. We establish that MinAres and cAr generate monotonic $\|x_k - x^\star\|$, $\|x_k - x^\star\|_A$ and $\|r_k\|$ when A is positive definite.

Keywords : MinAres, cAr, Minres, Cr, Lsmr, symmetric, singular, inconsistent, iterative method, Lanczos process, Krylov subspace, QR factorization, LQ factorization

---

Acknowledgements: Research of A. Montoison is supported by an FRQNT grant and an excellence scholarship of the IVADO institute. This work began while the first author was visiting ICME at Stanford University in Spring 2022. My thanks to Mike Saunders for making my stay such a pleasant one. Research of D. Orban is partially supported by an NSERC Discovery Grant.

1 Introduction

Suppose $A \in \mathbb{R}^{n\times n}$ is a large symmetric matrix for which matrix-vector products Av can be computed efficiently for any vector $v \in \mathbb{R}^n$. We present a Krylov subspace method called MinAres for computing a solution to the following problems:

$$\text{Symmetric linear systems:} \quad Ax = b, \tag{1}$$
$$\text{Symmetric least-squares problems:} \quad \min \|Ax - b\|, \tag{2}$$
$$\text{Symmetric nullspace problems:} \quad Ar = 0, \tag{3}$$
$$\text{Symmetric eigenvalue problems:} \quad Ar = \lambda r, \tag{4}$$
$$\text{Singular value problems for rectangular } B\text{:} \quad \begin{bmatrix} & B \\ B^T & \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \sigma \begin{bmatrix} u \\ v \end{bmatrix}. \tag{5}$$

If A is nonsingular, problems (1)–(2) have a unique solution $x^\star$. When A is singular, if b is not in the range of A then (1) has no solution; otherwise, (1)–(2) have an infinite number of solutions, and we seek the unique solution $x^\star$ that minimizes $\|x\|$. Whenever $x^\star$ exists, it solves the problem

$$\min \tfrac{1}{2}\|x\|^2 \quad \text{subject to} \quad A^2 x = Ab. \tag{6}$$

Let x_k be an approximation to $x^\star$ with residual $r_k = b - Ax_k$. If A were unsymmetric or rectangular, applicable solvers for (1)–(2) would be Lsqr [16] and Lsmr [4], which reduce $\|r_k\|$ and $\|A^T r_k\|$ respectively within the kth Krylov subspace $\mathcal{K}_k(A^TA, A^Tb)$ generated by the Golub-Kahan bidiagonalization on (A, b) [7].

For (1)–(5), we propose an algorithm MinAres that solves (6) by reducing $\|Ar_k\|$ within the kth Krylov subspace $\mathcal{K}_k(A, b)$ generated by the symmetric Lanczos process on (A, b) [11]. Thus when A is symmetric, MinAres minimizes the same quantity $\|Ar_k\|$ as Lsmr, but in different (more effective) subspaces, and it requires only one matrix-vector product Av per iteration, whereas Lsmr would need two.

Qualitatively, certain residual norms decrease smoothly for these iterative methods, but other norms are more erratic as they approach zero. It is ideal if stopping criteria involve the smooth quantities. For Lsqr and Lsmr on general (possibly rectangular) systems, $\|r_k\|$ decreases smoothly for both methods. We observe that while Lsqr is always ahead by construction, it is never by very much. Thus on consistent systems $Ax = b$, Lsqr may terminate slightly sooner. On inconsistent systems $Ax \approx b$, the comparison is more striking. $\|A^T r_k\|$ decreases erratically for Lsqr but smoothly for Lsmr, and there is usually a significance difference between the two. Thus Lsmr may terminate significantly sooner [4].

Similarly for Minres [15] and MinAres, $\|r_k\|$ decreases smoothly for both methods, and on consistent symmetric systems $Ax = b$, Minres may have a small advantage. On inconsistent symmetric systems $Ax \approx b$, $\|Ar_k\|$ decreases erratically for Minres and its variant Minres-qlp [2] but smoothly for MinAres, and there is usually a significant difference between them. Thus MinAres may terminate sooner.

We introduce cAr, a new conjugate direction method similar to Cg and Cr and equivalent to MinAres when A is SPD. We prove that $\|r_k\|$, $\|x_k - x^\star\|$ and $\|x_k - x^\star\|_A$ decrease monotonically for cAr and hence MinAres when A is positive definite.

1.1 Notation

A symmetric positive definite matrix is said to be SPD. For a vector v_k, $\|v_k\|$ denotes the Euclidean norm of v_k, and for an SPD matrix A, the A-norm of v_k is $\|v_k\|_A^2 = v^TAv$. For a matrix V_k, $\|V_k\|$ may be any norm. Vector e_j is the jth column of an identity matrix I_k of size dictated by the context. An approximate solution x_k has residual $r_k = b - Ax_k$, and $x^\star$ is the unique solution of

$Ax = b$ if A is nonsingular, or the minimum-norm solution of $A^2x = Ab$ otherwise. $\mathcal{K}_k(A, b)$ is the *Krylov subspace* $\{b, Ab, \dots, A^{k-1}b\}$. We abusively write $z = (\zeta_1, \dots, \zeta_n)$ to represent the column vector $z = \begin{bmatrix}\zeta_1 & \dots & \zeta_n\end{bmatrix}^T$. If H is SPD and $\{d_1, \dots, d_k\}$ is a set of non-zero vectors, the vectors are H-conjugate if $d_i^T H d_j = 0$ for $i \neq j$. If $H = I$, conjugacy is equivalent to the usual notion of orthogonality.

2 Applications

2.1 Null vector, eigenvector, and singular value problems

Given a symmetric A and nonzero b, MinAres solves $A^2x = Ab$ even if A is singular. If b is random and A is singular, $r = b - Ax$ is unlikely to be zero, but it will be a nonzero nullvector of A because $Ar = 0$.

If an eigenvalue λ of A is known, we can use it as a shift in the Lanczos process with a random starting vector b to find a null vector r such that $(A - \lambda I)r = 0$. Then r is an eigenvector because $Ar = \lambda r$. MinAres is effectively implementing the inverse power method [8, 18] to obtain the eigenvector in one iteration. If λ is approximate, MinAres can implement Rayleigh quotient iteration [8, 18] to obtain increasingly accurate eigenpair estimates.

Similarly, if a singular value σ is known for a rectangular matrix B, the singular value problem $Bv = \sigma u$, $B^T u = \sigma v$ may be reformulated as a null vector problem or eigenvalue problem:

$$\left(\begin{bmatrix} & B \\ B^T & \end{bmatrix} - \sigma I \right) \begin{bmatrix} u \\ v \end{bmatrix} = 0 \quad \Longleftrightarrow \quad \begin{bmatrix} & B \\ B^T & \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \sigma \begin{bmatrix} u \\ v \end{bmatrix},$$

for which MinAres may be used to implement inverse iteration or Rayleigh quotient iteration (although an algorithm based on the Golub-Kahan bidiagonalization of B would be preferable).

2.2 Singular systems with semi-positive definite matrices

Inconsistent (singular) symmetric systems could arise from discretized semidefinite Neumann boundary value problems [10, sect. 4]. Measurement errors will be random, so b is unlikely to be in the range of singular A.

Another potential application is large, singular, symmetric, indefinite Toeplitz least-squares problems as described in [6, sec. 5]. Rank-deficient Toeplitz matrices arise in image reconstruction and system identification problems. In both cases, A is a semi-positive definite matrix and MinAres is a suitable solver.

3 Symmetric systems

With A symmetric and starting vector b, we make use of the symmetric Lanczos process [11] of Algorithm 1. After k iterations the situation may be summarized as

$$AV_k = V_k T_k + \beta_{k+1} v_{k+1} e_k^T = V_{k+1} T_{k+1,k}, \tag{7a}$$

$$V_k^T V_k = I_k, \tag{7b}$$

where

$$V_k := \begin{bmatrix} v_1 & \dots & v_k \end{bmatrix}, \qquad T_k = \begin{bmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_k \\ & & \beta_k & \alpha_k \end{bmatrix}, \qquad T_{k+1,k} = \begin{bmatrix} T_k \\ \beta_{k+1} e_k^T \end{bmatrix}.$$

Algorithm 1 Lanczos process

```
Require: A, b
v_0 = 0
β_1 v_1 = b                                  β_1 > 0 so that ||v_1|| = 1
for k = 1, 2, ... do
    q_k = A v_k − β_k v_{k−1}
    α_k = v_k^T q_k
    q_k = q_k − α_k v_k
    β_{k+1} = ||q_k||
    if β_{k+1} = 0 then
        ℓ = k;  return ℓ
    else
        v_{k+1} = q_k / β_{k+1}              β_{k+1} > 0 so that ||v_{k+1}|| = 1
    end if
end for
```

In exact arithmetic, V_k is an orthonormal basis of $\mathcal{K}_k(A,b)$. The Lanczos process terminates after $\ell \le n$ iterations when $\beta_{\ell+1} = 0$, and we then have $AV_\ell = V_\ell T_\ell$, where square T_ℓ is nonsingular if and only if $b \in \operatorname{range}(A)$ [2, sec. 2.1 property 4]. $T_{k+1,k}$ has full column rank k for all $k < \ell$ [2, sec. 2.1 property 2] and the rank of T_ℓ is ℓ or $\ell - 1$ but no less (because the first $\ell - 1$ columns of T_ℓ are independent).

In finite arithmetic, (7a) holds to machine precision. Reorthogonalization would be needed for (7b) to hold accurately, but it is enough to note that we always have $\|V_k\| = O(1)$.

3.1 Cg, Symmlq, Minres, MinAres

As with Cg [9], Symmlq [15], and Minres [15], the goal of MinAres is to solve symmetric problems $Ax \approx b$. All methods define an approximate solution $x_k = V_k y_k$ at iteration k (where y_k is different for each method). MinAres chooses y_k to minimize $\|Ar_k\|$ in $\mathcal{K}_k(A,b)$, so that $\|Ar_k\|$ is monotonically decreasing towards zero. MinAres is therefore well suited to singular inconsistent symmetric systems. This case is difficult for the other methods because $\|x_k - x^\star\|_A$, $\|x_k - x^\star\|$ and $\|r_k\|$ do not converge to zero and they are the quantities minimized respectively by Cg, Symmlq, and both Minres and Minres-qlp.

4 Derivation of MinAres

4.1 Subproblems of MinAres

From Algorithm 1 we have $Ab = \beta_1\alpha_1 v_1 + \beta_1\beta_2 v_2$ because $\beta_2 v_2 = Av_1 - \alpha_1 v_1$. Hence

$$\begin{aligned}
Ar_k &= A(b - AV_k y_k) \\
&= Ab - AV_{k+1}T_{k+1,k}y_k \\
&= \beta_1\alpha_1 v_1 + \beta_1\beta_2 v_2 - V_{k+2}T_{k+2,k+1}T_{k+1,k}y_k \\
&= V_{k+2}(\beta_1\alpha_1 e_1 + \beta_1\beta_2 e_2 - T_{k+2,k+1}T_{k+1,k}y_k), \quad k \le \ell - 2, && \text{(8a)}\\
Ar_{\ell-1} &= V_\ell(\beta_1\alpha_1 e_1 + \beta_1\beta_2 e_2 - T_\ell T_{\ell,\ell-1}y_{\ell-1}), && \text{(8b)}\\
Ar_\ell &= V_\ell(\beta_1\alpha_1 e_1 + \beta_1\beta_2 e_2 - T_\ell^2 y_\ell). && \text{(8c)}
\end{aligned}$$

Theoretically, V_k has orthonormal columns ($1 \le k \le \ell$), so that $\|x_k\| = \|y_k\|$ and $\|Ar_k\|$ is minimized with $\|x_k\|$ of minimal norm if we define y_k as the unique solution of the following subproblems:

$$\underset{y_k \in \mathbb{R}^k}{\text{minimize}} \quad \|T_{k+2,k+1}T_{k+1,k}y_k - \beta_1\alpha_1 e_1 - \beta_1\beta_2 e_2\|, \quad k \le \ell - 2, \tag{9a}$$

$$\underset{y_{\ell-1} \in \mathbb{R}^{\ell-1}}{\text{minimize}} \quad \|T_\ell T_{\ell,\ell-1}y_{\ell-1} - \beta_1\alpha_1 e_1 - \beta_1\beta_2 e_2\|, \tag{9b}$$

$$\underset{y_\ell \in \mathbb{R}^\ell}{\text{minimize}} \quad \|y_\ell\|^2 \quad \text{subject to} \quad T_\ell^2 y_\ell = \beta_1\alpha_1 e_1 + \beta_1\beta_2 e_2. \tag{9c}$$

We define y_k from these subproblems even though V_k does not remain orthonormal numerically. In practice, we expect $\|Ar_k\| \le \|Ar_{k-1}\|$ unless k becomes too large.

To be sure that the subproblems have unique solutions, we need to verify that $T_{k+2,k+1}T_{k+1,k}$ has rank k ($k \le \ell - 2$), $T_\ell T_{\ell,\ell-1}$ has rank $\ell - 1$, and $T_\ell^2 y_\ell = \beta_1\alpha_1 e_1 + \beta_1\beta_2 e_2$ is consistent even if T_ℓ is singular. These results are proved in Theorem 1, Theorem 2 and Theorem 3.

Theorem 1. *For $k \le \ell - 2$, $T_{k+2,k+1}T_{k+1,k}$ has rank k.*

See proof on page 14.

Theorem 2. *$T_\ell T_{\ell,\ell-1}$ has rank $\ell - 1$.*

See proof on page 14.

Theorem 3. *$T_\ell^2 y_\ell = \beta_1\alpha_1 e_1 + \beta_1\beta_2 e_2$ is consistent even if T_ℓ is singular.*

See proof on page 14.

From (8c) and Theorem 3, $Ar_\ell = V_\ell(T_\ell^2 y_\ell - \beta_1\alpha_1 e_1 - \beta_1\beta_2 e_2) = 0$. Hence with definition (9c) we can conclude that x_ℓ is the solution $x^\star$ of (6).

4.2 QR factorization of T_k

To solve (9), we first need the QR factorization used by MINRES:

$$T_{k+1,k} = Q_k \begin{bmatrix} R_k \\ 0 \end{bmatrix}, \quad R_k = \begin{bmatrix} \lambda_1 & \gamma_1 & \varepsilon_1 & & \\ & \lambda_2 & \gamma_2 & \ddots & \\ & & \lambda_3 & \ddots & \varepsilon_{k-2} \\ & & & \ddots & \gamma_{k-1} \\ & & & & \lambda_k \end{bmatrix}, \tag{10}$$

where $Q_k^T = Q_{k+1,k} \dots Q_{3,2}Q_{2,1}$ is an orthogonal matrix defined as a product of 2×2 reflections with the structure

$$Q_{i+1,i} = \begin{array}{c} \\ 1 \\ \vdots \\ i-1 \\ i \\ i+1 \\ i+2 \\ \vdots \\ k \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & \dots & i-1 & i & i+1 & i+2 & \dots & k \end{array} \\ \begin{bmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & c_i & s_i & & & \\ & & & s_i & -c_i & & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{bmatrix} \end{array}.$$

If we initialize $Q_0 := I$, $\bar\lambda_1 := \alpha_1$, $\bar\gamma_1 := \beta_2$, individual factorization steps may be represented as an application of $Q_{k+1,k}$ to $Q_{k-1}^T T_{k+1,k}$:

$$\begin{array}{c} \\ k \\ k+1 \end{array} \begin{array}{c} \begin{array}{cc} k & k+1 \end{array} \\ \begin{bmatrix} c_k & s_k \\ s_k & -c_k \end{bmatrix} \end{array} \begin{array}{c} \begin{array}{ccc} k & k+1 & k+2 \end{array} \\ \left[\begin{array}{c|cc} \bar\lambda_k & \bar\gamma_k & 0 \\ \beta_{k+1} & \alpha_{k+1} & \beta_{k+2} \end{array}\right] \end{array} = \begin{array}{c} \begin{array}{ccc} k & k+1 & k+2 \end{array} \\ \left[\begin{array}{c|cc} \lambda_k & \gamma_k & \varepsilon_k \\ 0 & \bar\lambda_{k+1} & \bar\gamma_{k+1} \end{array}\right] \end{array}.$$

The reflection $Q_{k+1,k}$ zeroes β_{k+1} on the subdiagonal of $T_{k+1,k}$ and affects three columns and two rows. It is defined by

$$\lambda_k = \sqrt{\bar\lambda_k^2 + \beta_{k+1}^2}, \quad c_k = \bar\lambda_k/\lambda_k, \quad s_k = \beta_{k+1}/\lambda_k, \tag{11}$$

and yields the following recursion for $k \ge 1$:

$$\gamma_k = c_k\bar\gamma_k + s_k\alpha_{k+1}, \tag{12a}$$
$$\bar\lambda_{k+1} = s_k\bar\gamma_k - c_k\alpha_{k+1}, \tag{12b}$$
$$\varepsilon_k = s_k\beta_{k+2}, \tag{12c}$$
$$\bar\gamma_{k+1} = -c_k\beta_{k+2}. \tag{12d}$$

4.3 Definition of N_k

Let us define

$$N_k := T_{k+2,k+1} Q_k \begin{bmatrix} I_k \\ 0 \end{bmatrix}, \quad \text{where} \quad N_k R_k = T_{k+2,k+1} T_{k+1,k}, \quad k \le \ell - 2, \tag{13a}$$

$$N_{\ell-1} := T_{\ell,\ell-1} Q_{\ell-1} \begin{bmatrix} I_{\ell-1} \\ 0 \end{bmatrix}, \quad \text{where} \quad N_{\ell-1} R_{\ell-1} = T_\ell T_{\ell,\ell-1}, \tag{13b}$$

$$N_\ell := T_\ell Q_\ell, \qquad \text{where} \quad N_\ell R_\ell = T_\ell^2. \tag{13c}$$

Because $Q_k = Q_{2,1} Q_{3,2} \dots Q_{k+1,k}$, we have

$$e_k^T Q_k = e_k^T Q_{k,k-1} Q_{k+1,k} = s_{k-1} e_{k-1}^T - c_{k-1} c_k e_k^T - c_{k-1} s_k e_{k+1}^T, \tag{14a}$$

$$e_{k+1}^T Q_k = e_{k+1}^T Q_{k+1,k} = s_k e_k^T - c_k e_{k+1}^T. \tag{14b}$$

Moreover, $T_{k+2,k+1} = \begin{bmatrix} T_{k+1,k}^T \\ \beta_{k+1} e_k^T + \alpha_{k+1} e_{k+1}^T \\ \beta_{k+2} e_{k+1}^T \end{bmatrix}$ and the product $T_{k+2,k+1} Q_k$ can be determined in three parts. From (10), $T_{k+1,k}^T Q_k = \left(Q_k^T T_{k+1,k} \right)^T = \begin{bmatrix} R_k^T & 0 \end{bmatrix}$, and from (14) we have

$$\begin{aligned}
(\beta_{k+1} e_k^T + \alpha_{k+1} e_{k+1}^T) Q_k &= \beta_{k+1} s_{k-1} e_{k-1}^T + (\alpha_{k+1} s_k - \beta_{k+1} c_{k-1} s_k) e_k^T \\
&\qquad - (\alpha_{k+1} c_k + \beta_{k+1} c_{k-1} s_k) e_{k+1}^T \\
&= \varepsilon_{k-1} e_{k-1}^T + \gamma_k e_k^T - (\alpha_{k+1} c_k + \beta_{k+1} c_{k-1} s_k) e_{k+1}^T, \\
\beta_{k+2} e_{k+1}^T Q_k &= s_k \beta_{k+2} e_k^T - c_k \beta_{k+2} e_{k+1}^T \\
&= \varepsilon_k e_k^T - c_k \beta_{k+2} e_{k+1}^T.
\end{aligned}$$

Thus, for $k \le \ell - 2$ we obtain

$$N_k = \begin{bmatrix} R_k^T \\ \varepsilon_{k-1} e_{k-1}^T + \gamma_k e_k^T \\ \varepsilon_k e_k^T \end{bmatrix}, \quad N_{\ell-1} = \begin{bmatrix} R_{\ell-1}^T \\ \varepsilon_{\ell-1} e_{\ell-1}^T + \gamma_\ell e_\ell^T \end{bmatrix}, \quad N_\ell = R_\ell^T. \tag{15}$$

4.4 QR factorization of N_k

$$N_k = \widetilde{Q}_k \begin{bmatrix} U_k \\ 0 \end{bmatrix}, \quad U_k = \begin{bmatrix} \mu_1 & \phi_1 & \rho_1 & & \\ & \mu_2 & \phi_2 & \ddots & \\ & & \mu_3 & \ddots & \rho_{k-2} \\ & & & \ddots & \phi_{k-1} \\ & & & & \mu_k \end{bmatrix}, \tag{16}$$

where $\widetilde{Q}_k^T = \widetilde{Q}_{k+2,k} \widetilde{Q}_{k+1,k} \dots \widetilde{Q}_{3,1} \widetilde{Q}_{2,1}$ for $k \le \ell - 2$, and $\widetilde{Q}_\ell^T = \widetilde{Q}_{\ell-1}^T = \widetilde{Q}_{\ell,\ell-1} \widetilde{Q}_{\ell-2}^T$ are orthogonal matrices defined as a product of reflections. If we initialize $\bar{\mu}_1 := \lambda_1$, $\hat{\gamma}_1 := \gamma_1$ and $\hat{\lambda}_2 := \lambda_2$, individual factorization steps may be represented as an application of $\widetilde{Q}_{k+1,k}$ to $\widetilde{Q}_{k-1}^T N_k$:

$$\begin{matrix} k \\ k{+}1 \\ k{+}2 \end{matrix} \overset{\begin{matrix} k & k+1 & k+2 \end{matrix}}{\begin{bmatrix} \tilde{c}_{2k-1} & \tilde{s}_{2k-1} & \\ \tilde{s}_{2k-1} & -\tilde{c}_{2k-1} & \\ & & 1 \end{bmatrix}} \overset{\begin{matrix} k & k+1 & k+2 \end{matrix}}{\left[\begin{array}{c|cc} \bar{\mu}_k & & \\ \hat{\gamma}_k & \hat{\lambda}_{k+1} & \\ \varepsilon_k & \gamma_{k+1} & \lambda_{k+2} \end{array}\right]} = \overset{\begin{matrix} k & k+1 & k+2 \end{matrix}}{\left[\begin{array}{c|cc} \bar{\bar{\mu}}_k & \bar{\phi}_k & \\ & \bar{\mu}_{k+1} & \\ \varepsilon_k & \gamma_{k+1} & \lambda_{k+2} \end{array}\right]},$$

followed by an application of $Q_{k+2,k}$ to the result:

$$\begin{matrix} k \\ k+1 \\ k+2 \end{matrix}\overset{\begin{matrix} k & k+1 & k+2 \end{matrix}}{\begin{bmatrix} \tilde{c}_{2k} & & \tilde{s}_{2k} \\ & 1 & \\ \tilde{s}_{2k} & & -\tilde{c}_{2k} \end{bmatrix}} \overset{\begin{matrix} k & k+1 & k+2 \end{matrix}}{\left[\begin{array}{c|cc} \bar{\bar{\mu}}_k & \bar{\phi}_k & \\ & \bar{\mu}_{k+1} & \\ \varepsilon_k & \gamma_{k+1} & \lambda_{k+2} \end{array}\right]} = \overset{\begin{matrix} k & k+1 & k+2 \end{matrix}}{\left[\begin{array}{c|cc} \mu_k & \phi_k & \rho_k \\ & \bar{\mu}_{k+1} & \\ & \hat{\gamma}_{k+1} & \hat{\lambda}_{k+2} \end{array}\right]}.$$

The reflections $\widetilde{Q}_{k+1,k}$ and $\widetilde{Q}_{k+2,k}$ zero γ_k and ε_k on the subdiagonals of N_k:

$$\bar{\bar{\mu}}_k = \sqrt{\bar{\mu}_k^2 + \hat{\gamma}_k^2}, \quad \tilde{c}_{2k-1} = \bar{\mu}_k/\bar{\bar{\mu}}_k, \quad \tilde{s}_{2k-1} = \hat{\gamma}_k/\bar{\bar{\mu}}_k, \quad k \le \ell - 1, \tag{17a}$$

$$\mu_k = \sqrt{\bar{\bar{\mu}}_k^2 + \varepsilon_k^2}, \quad \tilde{c}_{2k} = \bar{\bar{\mu}}_k/\mu_k, \quad \tilde{s}_{2k} = \varepsilon_k/\mu_k, \quad k \le \ell - 2, \tag{17b}$$

and they yield the recursion

$$\bar{\phi}_k = \tilde{s}_{2k-1}\hat{\lambda}_{k+1}, \qquad 1 \le k \le \ell - 1, \tag{18a}$$

$$\bar{\mu}_{k+1} = -\tilde{c}_{2k-1}\hat{\lambda}_{k+1}, \qquad 1 \le k \le \ell - 1, \tag{18b}$$

$$\phi_k = \tilde{c}_{2k}\bar{\phi}_k + \tilde{s}_{2k}\gamma_{k+1}, \quad 1 \le k \le \ell - 2, \tag{18c}$$

$$\hat{\gamma}_{k+1} = \tilde{s}_{2k}\bar{\phi}_k - \tilde{c}_{2k}\gamma_{k+1}, \quad 1 \le k \le \ell - 2, \tag{18d}$$

$$\rho_k = \tilde{s}_{2k}\lambda_{k+2}, \qquad 1 \le k \le \ell - 2, \tag{18e}$$

$$\hat{\lambda}_{k+2} = -\tilde{c}_{2k}\lambda_{k+2}, \qquad 1 \le k \le \ell - 2, \tag{18f}$$

$$\mu_{\ell-1} = \bar{\bar{\mu}}_{\ell-1}, \tag{18g}$$

$$\phi_{\ell-1} = \bar{\phi}_{\ell-1}, \tag{18h}$$

$$\mu_\ell = \bar{\mu}_\ell. \tag{18i}$$

From (8) and (16) we have

$$\|Ar_k\| = \|N_k R_k y_k - \beta_1\alpha_1 e_1 - \beta_1\beta_2 e_2\| = \left\| \begin{bmatrix} U_k \\ 0 \end{bmatrix} R_k y_k - \bar{z}_k \right\|, \tag{19}$$

where $\bar{z}_k := \widetilde{Q}_k^T(\beta_1\alpha_1 e_1 + \beta_1\beta_2 e_2) = (z_k, \bar{\bar{\zeta}}_{k+1}, \bar{\zeta}_{k+2})$, $k \le \ell - 2$, $z_k = (\zeta_1, \dots, \zeta_k)$ represents the first k components of $\bar{z}_k$, and the recurrence starts with $\bar{z}_0 := (\bar{\bar{\zeta}}_1, \bar{\zeta}_2) = (\beta_1\alpha_1, \beta_1\beta_2)$. We can determine $\bar{z}_k$ from $\bar{z}_{k-1}$ because $\bar{z}_k = \widetilde{Q}_{k+2,k}\widetilde{Q}_{k+1,k}(\bar{z}_{k-1}, 0)$ for $k \le \ell - 2$:

$$\begin{matrix} k \\ k+1 \\ k+2 \end{matrix}\overset{\begin{matrix} k & k+1 & k+2 \end{matrix}}{\begin{bmatrix} \tilde{c}_{2k} & & \tilde{s}_{2k} \\ & 1 & \\ \tilde{s}_{2k} & & -\tilde{c}_{2k} \end{bmatrix}} \overset{\begin{matrix} k & k+1 & k+2 \end{matrix}}{\begin{bmatrix} \tilde{c}_{2k-1} & \tilde{s}_{2k-1} & \\ \tilde{s}_{2k-1} & -\tilde{c}_{2k-1} & \\ & & 1 \end{bmatrix}} \begin{bmatrix} \bar{\bar{\zeta}}_k \\ \bar{\zeta}_{k+1} \\ 0 \end{bmatrix} = \begin{bmatrix} \zeta_k \\ \bar{\bar{\zeta}}_{k+1} \\ \bar{\zeta}_{k+2} \end{bmatrix},$$

and $z_\ell = z_{\ell-1} = \widetilde{Q}_{\ell,\ell-1}\bar{z}_{\ell-2}$. The elements are updated according to

$$\mathring{\zeta}_k = \tilde{c}_{2k-1}\bar{\bar{\zeta}}_k + \tilde{s}_{2k-1}\bar{\zeta}_{k+1}, \quad k \le \ell - 1, \tag{20a}$$

$$\bar{\bar{\zeta}}_{k+1} = \tilde{s}_{2k-1}\bar{\bar{\zeta}}_k - \tilde{c}_{2k-1}\bar{\zeta}_{k+1}, \quad k \le \ell - 1, \tag{20b}$$

$$\zeta_k = \tilde{c}_{2k}\mathring{\zeta}_k, \qquad k \le \ell - 2, \tag{20c}$$

$$\bar{\zeta}_{k+2} = \tilde{s}_{2k}\mathring{\zeta}_k, \qquad k \le \ell - 2, \tag{20d}$$

$$\zeta_{\ell-1} = \mathring{\zeta}_{\ell-1}, \tag{20e}$$

$$\zeta_\ell = \bar{\bar{\zeta}}_\ell. \tag{20f}$$

For $k \le \ell - 1$, U_k and R_k are nonsingular, and from (19), $\|Ar_k\|$ is minimized when $U_k R_k y_k = z_k$, giving

$$\|Ar_k\| = \sqrt{\bar{\bar{\zeta}}_{k+1}^2 + \bar{\zeta}_{k+2}^2}, \quad k \le \ell - 2, \quad \|Ar_{\ell-1}\| = |\zeta_\ell|. \tag{21}$$

4.5 Computation of x_k

Suppose R_k and U_k are nonsingular. If we were to update x_k directly from $x_k = V_k y_k$, all components of y_k would have to be recomputed because of the backward substitutions required to solve $U_k R_k y_k = z_k$, which would require us to store V_k entirely. To avoid such drawbacks, we employ the strategy of Paige and Saunders [15]. Thus, we define W_k and D_k by the lower triangular systems $R_k^T W_k^T = V_k^T$ and $U_k^T D_k^T = W_k^T$. Then

$$x_k = V_k y_k = W_k R_k y_k = D_k U_k R_k y_k = D_k z_k. \tag{22}$$

The columns of W_k and D_k are obtained from the recursions

$$\begin{aligned}
w_1 &= v_1/\lambda_1, \qquad w_2 = (v_2 - \gamma_1 w_1)/\lambda_2, \\
w_k &= (v_k - \gamma_{k-1} w_{k-1} - \varepsilon_{k-2} w_{k-2})/\lambda_k, \quad k \geq 3, \\
d_1 &= w_1/\mu_1, \qquad d_2 = (w_2 - \phi_1 d_1)/\mu_2, \\
d_k &= (w_k - \phi_{k-1} d_{k-1} - \rho_{k-2} d_{k-2})/\mu_k, \quad k \geq 3,
\end{aligned}$$

and the solution $x_k = D_k z_k$ may be updated efficiently via $x_0 = 0$ and

$$x_k = x_{k-1} + \zeta_k d_k. \tag{23}$$

This is possible for all $k \leq \ell$ if $Ax = b$ is consistent, and $k \leq \ell - 1$ otherwise. If $Ax = b$ is consistent, from Theorem 4, the final MinAres iterate x_ℓ satisfies $r_\ell = 0$ and is the minimum-length solution. If $Ax = b$ is inconsistent, from Theorem 5, $Ar_{\ell-1} = 0$. We obtain a solution x that satisfies $A^2 x = Ab$ in both cases.

Theorem 4. *If $b \in range(A)$, the final* MinAres *iterate x_ℓ is the minimum-length solution of $Ax = b$ (and $r_\ell = b - Ax_\ell = 0$).*

See proof on page 14.

Theorem 5. *If $Ax = b$ is inconsistent, $\zeta_\ell = 0$ and $Ar_{\ell-1} = 0$.*

See proof on page 15.

If the minimum-norm solution is not required, such as problems (3)–(5), we can stop with $x_{\ell-1}$ and avoid the computation of $x_\ell = x^\star$. We can also stop with $x_{\ell-1}$ if a preconditioner is used because the minimum-norm solution is determined in a non-Euclidean norm.

We summarize the complete procedure as Algorithm 2.

5 Stopping rules

The end of Algorithm 2 shows how $\|r_k\|$ and $\|Ar_k\|$ are estimated. They are needed for use within stopping rules. The required norm estimates are derived next.

5.1 Estimating $\|r_k\|$

To compute $\|r_k\|$, we need an LQ factorization

$$U_k = \hat{L}_k \hat{P}_k, \quad \hat{L}_k = \begin{bmatrix}
\psi_1 & & & & & & \\
\theta_1 & \psi_2 & & & & & \\
\omega_1 & \theta_2 & \psi_3 & & & & \\
& \ddots & \ddots & \ddots & & & \\
& & \ddots & \ddots & \psi_{k-2} & & \\
& & & \ddots & \theta_{k-2} & \bar{\bar{\psi}}_{k-1} & \\
& & & & \omega_{k-2} & \bar{\theta}_{k-1} & \bar{\psi}_k
\end{bmatrix}, \tag{24}$$

where $\hat{P}_1^T = I$, $\hat{P}_2^T = \hat{P}_{1,2}$, and $\hat{P}_k^T = \hat{P}_{k-1}^T \hat{P}_{k-2,k} \hat{P}_{k-1,k}$ $(k \geq 3)$ are orthogonal. Note that $\hat{L}_k$ is the L factor of a QLP decomposition of N_k. If we initialize $\bar{\psi}_1 := \mu_1$, $\hat{P}_{1,2}$ is defined to zero ϕ_1:

$$\begin{bmatrix} \bar{\psi}_1 & \phi_1 \\ & \mu_2 \end{bmatrix} \begin{bmatrix} \hat{c}_1 & \hat{s}_1 \\ \hat{s}_1 & -\hat{c}_1 \end{bmatrix} = \begin{bmatrix} \bar{\bar{\psi}}_1 & \\ \bar{\theta}_1 & \bar{\psi}_2 \end{bmatrix},$$

Algorithm 2 MinAres

Require: A, b, $\epsilon_r > 0$, $\epsilon_{Ar} > 0$, $k_{\max} > 0$

$k = 0, \quad x_0 = 0$

$w_{-1} = w_0 = 0, \quad d_{-1} = d_0 = 0$

$\varepsilon_{-1} = \varepsilon_0 = \gamma_0 = 0, \quad \rho_{-1} = \rho_0 = \phi_0 = 0$

$\beta_1 v_1 = b, \quad q_1 = A v_1, \quad \alpha_1 = v_1^T q_1$

$q_1 = q_1 - \alpha_1 v_1, \quad \beta_2 v_2 = q_1$

$\bar{\bar{\zeta}}_1 = \beta_1 \alpha_1, \quad \bar{\zeta}_2 = \beta_1 \beta_2$

$\bar{\chi}_1 = \beta_1, \quad \bar{\lambda}_1 = \alpha_1, \quad \bar{\gamma}_1 = \beta_2$

$\|r_0\| = \bar{\chi}_1, \quad \|A r_0\| = (\bar{\bar{\zeta}}_1^2 + \bar{\zeta}_2^2)^{\frac{1}{2}}$

while $\|r_k\| > \epsilon_r$ **and** $\|A r_k\| > \epsilon_{Ar}$ **and** $k \leq k_{\max}$ **do**

  $k \leftarrow k + 1$

  $q_{k+1} = A v_{k+1} - \beta_{k+1} v_k, \quad \alpha_{k+1} = v_{k+1}^T q_{k+1}$

  $q_{k+1} = q_{k+1} - \alpha_{k+1} v_{k+1}, \quad \beta_{k+2} v_{k+2} = q_{k+1}$

  $\lambda_k = (\bar{\lambda}_k^2 + \beta_{k+1}^2)^{\frac{1}{2}}, \quad c_k = \bar{\lambda}_k / \lambda_k, \quad s_k = \beta_{k+1} / \lambda_k$

  $\gamma_k = c_k \bar{\gamma}_k + s_k \alpha_{k+1}, \quad \varepsilon_k = s_k \beta_{k+2}$

  $\bar{\lambda}_{k+1} = s_k \bar{\gamma}_k - c_k \alpha_{k+1}, \quad \bar{\gamma}_{k+1} = -c_k \beta_{k+2}$

  if $k == 1$ **then**

    $\bar{\mu}_k = \lambda_k, \quad \hat{\gamma}_k = \gamma_k$

  else

    if $k == 2$ **then**

      $\hat{\lambda}_k = \lambda_k$

    else

      $\rho_{k-2} = \tilde{s}_{2k-4} \lambda_k, \quad \hat{\lambda}_k = -\tilde{c}_{2k-4} \lambda_k$

    end if

    $\bar{\phi}_{k-1} = \tilde{s}_{2k-3} \hat{\lambda}_k, \quad \phi_{k-1} = \tilde{c}_{2k-2} \bar{\phi}_{k-1} + \tilde{s}_{2k-2} \gamma_k$

    $\bar{\mu}_k = -\tilde{c}_{2k-3} \hat{\lambda}_k, \quad \hat{\gamma}_k = \tilde{s}_{2k-2} \bar{\phi}_{k-1} - \tilde{c}_{2k-2} \gamma_k$

  end if

  $\bar{\bar{\mu}}_k = (\bar{\mu}_k^2 + \hat{\gamma}_k^2)^{\frac{1}{2}}, \quad \tilde{c}_{2k-1} = \bar{\mu}_k / \bar{\bar{\mu}}_k, \quad \tilde{s}_{2k-1} = \hat{\gamma}_k / \bar{\bar{\mu}}_k$

  $\mu_k = (\bar{\bar{\mu}}_k^2 + \varepsilon_k^2)^{\frac{1}{2}}, \quad \tilde{c}_{2k} = \bar{\bar{\mu}}_k / \mu_k, \quad \tilde{s}_{2k} = \varepsilon_k / \mu_k$

  $\mathring{\zeta}_k = \tilde{c}_{2k-1} \bar{\bar{\zeta}}_k + \tilde{s}_{2k-1} \bar{\zeta}_{k+1}, \quad \zeta_k = \tilde{c}_{2k} \mathring{\zeta}_k$

  $\bar{\bar{\zeta}}_{k+1} = \tilde{s}_{2k-1} \bar{\bar{\zeta}}_k - \tilde{c}_{2k-1} \bar{\zeta}_{k+1}, \quad \bar{\zeta}_{k+2} = \tilde{s}_{2k} \mathring{\zeta}_k$

  $w_k = (v_k - \gamma_{k-1} w_{k-1} - \varepsilon_{k-2} w_{k-2}) / \lambda_k$

  $d_k = (w_k - \phi_{k-1} d_{k-1} - \rho_{k-2} d_{k-2}) / \mu_k$

  $x_k = x_{k-1} + \zeta_k d_k$

  $\|A r_k\| = (\bar{\bar{\zeta}}_{k+1}^2 + \bar{\zeta}_{k+2}^2)^{\frac{1}{2}}$

  $\chi_k = c_k \bar{\chi}_k, \quad \bar{\chi}_{k+1} = s_k \bar{\chi}_k$

  if $k == 1$ **then**

    $\bar{\psi}_k = \mu_k, \quad \bar{\bar{\pi}}_{k-1} = 0, \quad \bar{\pi}_k = \chi_k$

    $\xi_k = \zeta_k, \quad \bar{\bar{\tau}}_{k-1} = 0, \quad \bar{\tau}_k = \xi_k / \bar{\psi}_k$

  else if $k == 2$ **then**

    $\bar{\bar{\psi}}_{k-1} = (\bar{\psi}_{k-1}^2 + \phi_{k-1}^2)^{\frac{1}{2}}, \quad \hat{c}_{k-1} = \bar{\psi}_{k-1} / \bar{\bar{\psi}}_{k-1}, \quad \hat{s}_{k-1} = \phi_{k-1} / \bar{\bar{\psi}}_{k-1}$

    $\bar{\theta}_{k-1} = \hat{s}_{2k-3} \mu_k, \quad \bar{\psi}_k = -\hat{c}_{2k-3} \mu_k$

    $\bar{\bar{\pi}}_{k-1} = \hat{c}_{2k-3} \bar{\pi}_{k-1} + \hat{s}_{2k-3} \chi_k, \quad \bar{\pi}_k = \hat{s}_{2k-3} \bar{\pi}_{k-1} - \hat{c}_{2k-3} \chi_k$

    $\xi_k = \zeta_k, \quad \bar{\bar{\tau}}_{k-1} = \xi_{k-1} / \bar{\bar{\psi}}_{k-1}, \quad \bar{\tau}_k = (\xi_k - \bar{\theta}_{k-1} \bar{\tau}_{k-1}) / \bar{\psi}_k$

  else

    $\psi_{k-2} = (\bar{\bar{\psi}}_{k-2}^2 + \rho_{k-2}^2)^{\frac{1}{2}}, \quad \hat{c}_{2k-4} = \bar{\bar{\psi}}_{k-2} / \psi_{k-2}, \quad \hat{s}_{2k-4} = \rho_{k-2} / \psi_{k-2}$

    $\bar{\bar{\psi}}_{k-1} = (\bar{\psi}_{k-1}^2 + \delta_k^2)^{\frac{1}{2}}, \quad \hat{c}_{2k-3} = \bar{\psi}_{k-1} / \bar{\bar{\psi}}_{k-1}, \quad \hat{s}_{2k-3} = \delta_k / \bar{\bar{\psi}}_{k-1}$

    $\theta_{k-2} = \hat{c}_{2k-4} \bar{\theta}_{k-2} + \hat{s}_{2k-4} \phi_{k-1}, \quad \omega_{k-2} = \hat{s}_{2k-4} \mu_k$

    $\delta_k = \hat{s}_{2k-4} \bar{\theta}_{k-2} - \hat{c}_{2k-4} \phi_{k-1}, \quad \eta_k = -\hat{c}_{2k-4} \mu_k$

    $\bar{\theta}_{k-1} = \hat{s}_{2k-3} \eta_k, \quad \bar{\psi}_k = -\hat{c}_{2k-3} \eta_k, \quad \upsilon_k = \hat{s}_{2k-4} \bar{\bar{\pi}}_{k-2} - \hat{c}_{2k-4} \chi_k$

    $\bar{\bar{\pi}}_{k-1} = \hat{c}_{2k-3} \bar{\pi}_{k-1} + \hat{s}_{2k-3} \upsilon_k, \quad \bar{\pi}_k = \hat{s}_{2k-3} \bar{\pi}_{k-1} - \hat{c}_{2k-3} \upsilon_k$

    $\tau_{k-2} = \bar{\bar{\tau}}_{k-2} \bar{\bar{\psi}}_{k-2} / \psi_{k-2}, \quad \xi_k = \zeta_k - \omega_{k-2} \tau_{k-2}$

    $\bar{\tau}_{k-1} = (\xi_{k-1} - \theta_{k-2} \tau_{k-2}) / \bar{\bar{\psi}}_{k-1}, \quad \bar{\tau}_k = (\xi_k - \bar{\theta}_{k-1} \bar{\tau}_{k-1}) / \bar{\psi}_k$

  end if

  $\|r_k\| = ((\bar{\bar{\pi}}_{k-1} - \bar{\bar{\tau}}_{k-1})^2 + (\bar{\pi}_k - \bar{\tau}_k)^2 + \bar{\chi}_{k+1}^2)^{\frac{1}{2}}$

end while

where

$$\bar{\bar{\psi}}_1 = \sqrt{\bar{\psi}_1^2 + \phi_1^2}, \quad \hat{c}_1 = \bar{\psi}_1/\bar{\bar{\psi}}_1, \quad \hat{s}_1 = \phi_1/\bar{\bar{\psi}}_1, \quad \bar{\theta}_1 = \hat{s}_1\mu_2, \quad \bar{\psi}_2 = -\hat{c}_1\mu_2. \tag{25}$$

For $k \geq 3$, individual factorization steps may be represented as an application of $\hat{P}_{k-2,k}$ to $U_k\hat{P}_{k-1}^T$:

$$\begin{array}{c} \\ {\scriptstyle k-2} \\ {\scriptstyle k-1} \\ {\scriptstyle k} \end{array}
\begin{array}{c} \begin{array}{ccc} {\scriptstyle k-2} & {\scriptstyle k-1} & {\scriptstyle k} \end{array} \\ \begin{bmatrix} \bar{\bar{\psi}}_{k-2} & & \rho_{k-2} \\ \bar{\theta}_{k-2} & \bar{\psi}_{k-1} & \phi_{k-1} \\ & & \mu_k \end{bmatrix} \end{array}
\begin{array}{c} \begin{array}{ccc} {\scriptstyle k-2} & {\scriptstyle k-1} & {\scriptstyle k} \end{array} \\ \begin{bmatrix} \hat{c}_{2k-4} & & \hat{s}_{2k-4} \\ & 1 & \\ \hat{s}_{2k-4} & & -\hat{c}_{2k-4} \end{bmatrix} \end{array}
=
\begin{array}{c} \begin{array}{ccc} {\scriptstyle k-2} & {\scriptstyle k-1} & {\scriptstyle k} \end{array} \\ \begin{bmatrix} \psi_{k-2} & & \\ \theta_{k-2} & \bar{\psi}_{k-1} & \delta_k \\ \omega_{k-2} & & \eta_k \end{bmatrix} \end{array},$$

followed by an application of $\hat{P}_{k-1,k}$ to the result:

$$\begin{array}{c} \\ {\scriptstyle k-2} \\ {\scriptstyle k-1} \\ {\scriptstyle k} \end{array}
\begin{array}{c} \begin{array}{ccc} {\scriptstyle k-2} & {\scriptstyle k-1} & {\scriptstyle k} \end{array} \\ \begin{bmatrix} \psi_{k-2} & & \\ \theta_{k-2} & \bar{\psi}_{k-1} & \delta_k \\ \omega_{k-2} & & \eta_k \end{bmatrix} \end{array}
\begin{array}{c} \begin{array}{ccc} {\scriptstyle k-2} & {\scriptstyle k-1} & {\scriptstyle k} \end{array} \\ \begin{bmatrix} 1 & & \\ & \hat{c}_{2k-3} & \hat{s}_{2k-3} \\ & \hat{s}_{2k-3} & -\hat{c}_{2k-3} \end{bmatrix} \end{array}
=
\begin{array}{c} \begin{array}{ccc} {\scriptstyle k-2} & {\scriptstyle k-1} & {\scriptstyle k} \end{array} \\ \begin{bmatrix} \psi_{k-2} & & \\ \theta_{k-2} & \bar{\bar{\psi}}_{k-1} & \\ \omega_{k-2} & \bar{\theta}_{k-1} & \bar{\psi}_k \end{bmatrix} \end{array}.$$

The reflections $\hat{P}_{k-2,k}$ and $\hat{P}_{k-1,k}$ zero ρ_{k-2} and δ_k on the superdiagonals of U_k:

$$\psi_{k-2} = \sqrt{\bar{\bar{\psi}}_{k-2}^2 + \rho_{k-2}^2}, \quad \hat{c}_{2k-4} = \bar{\bar{\psi}}_{k-2}/\psi_{k-2}, \quad \hat{s}_{2k-4} = \rho_{k-2}/\psi_{k-2}, \tag{26a}$$

$$\bar{\bar{\psi}}_{k-1} = \sqrt{\bar{\psi}_{k-1}^2 + \delta_k^2}, \qquad \hat{c}_{2k-3} = \bar{\psi}_{k-1}/\bar{\bar{\psi}}_{k-1}, \quad \hat{s}_{2k-3} = \delta_k/\bar{\bar{\psi}}_{k-1}, \tag{26b}$$

and for $k \geq 3$ they yield the recursion

$$\theta_{k-2} = \hat{c}_{2k-4}\bar{\theta}_{k-2} + \hat{s}_{2k-4}\phi_{k-1}, \tag{27a}$$

$$\delta_k = \hat{s}_{2k-4}\bar{\theta}_{k-2} - \hat{c}_{2k-4}\phi_{k-1}, \tag{27b}$$

$$\omega_{k-2} = \hat{s}_{2k-4}\mu_k, \tag{27c}$$

$$\eta_k = -\hat{c}_{2k-4}\mu_k, \tag{27d}$$

$$\bar{\theta}_{k-1} = \hat{s}_{2k-3}\eta_k, \tag{27e}$$

$$\bar{\psi}_k = -\hat{c}_{2k-3}\eta_k. \tag{27f}$$

Assuming orthonormality of V_{k+1}, we have

$$\begin{aligned} \|r_k\| = \|\beta_1 e_1 - T_{k+1,k}y_k\| &= \left\| Q_k^T\beta_1 e_1 - \begin{bmatrix} R_k \\ 0 \end{bmatrix} y_k \right\| \\ &= \left\| \begin{bmatrix} \hat{P}_k & \\ & 1 \end{bmatrix} Q_k^T\beta_1 e_1 - \begin{bmatrix} \hat{P}_k R_k y_k \\ 0 \end{bmatrix} \right\| \\ &= \left\| p_{k+1} - \begin{bmatrix} t_k \\ 0 \end{bmatrix} \right\|, \end{aligned} \tag{28}$$

where

$$(\chi_1, \ldots, \chi_k, \bar{\chi}_{k+1}) := Q_k^T\beta_1 e_1, \tag{29a}$$

$$p_{k+1} := (\pi_1, \ldots, \pi_{k-2}, \bar{\bar{\pi}}_{k-1}, \bar{\pi}_k, \bar{\chi}_{k+1}) = \begin{bmatrix} \hat{P}_k & \\ & 1 \end{bmatrix} Q_k^T\beta_1 e_1, \tag{29b}$$

$$t_k := (\tau_1, \ldots, \tau_{k-2}, \bar{\bar{\tau}}_{k-1}, \bar{\tau}_k) \quad \text{solves} \quad \hat{L}_k t_k = z_k. \tag{29c}$$

The components of $Q_k^T\beta_1 e_1$ can be updated with the relations

$$\bar{\chi}_1 = \beta_1, \qquad \chi_k = c_k\bar{\chi}_k, \qquad \bar{\chi}_{k+1} = s_k\bar{\chi}_k, \tag{30}$$

the components of p_{k+1} are updated with

$$\bar{\pi}_1 = \chi_1, \tag{31a}$$

$$
\begin{aligned}
\upsilon_2 &= \chi_2, && (31b)\\
\pi_{k-2} &= \hat{c}_{2k-4}\bar{\bar{\pi}}_{k-2} + \hat{s}_{2k-4}\chi_k, \quad k \geq 3, && (31c)\\
\upsilon_k &= \hat{s}_{2k-4}\bar{\bar{\pi}}_{k-2} - \hat{c}_{2k-4}\chi_k, \quad k \geq 3, && (31d)\\
\bar{\bar{\pi}}_{k-1} &= \hat{c}_{2k-3}\bar{\pi}_{k-1} + \hat{s}_{2k-3}\upsilon_k, \quad k \geq 2, && (31e)\\
\bar{\pi}_k &= \hat{s}_{2k-3}\bar{\pi}_{k-1} - \hat{c}_{2k-3}\upsilon_k, \quad k \geq 2, && (31f)
\end{aligned}
$$

and with $\omega_{-1} = \omega_0 = \theta_0 = \bar{\theta}_0 = 0$ the components of t_k are updated with

$$
\begin{aligned}
\xi_k &= \zeta_k - \omega_{k-2}\tau_{k-2}, && (32a)\\
\bar{\tau}_k &= (\xi_k - \bar{\theta}_{k-1}\bar{\tau}_{k-1})/\bar{\psi}_k, && (32b)\\
\bar{\bar{\tau}}_k &= (\xi_k - \theta_{k-1}\tau_{k-1})/\bar{\bar{\psi}}_k, && (32c)\\
\tau_k &= \bar{\bar{\tau}}_k\bar{\bar{\psi}}_k/\psi_k. && (32d)
\end{aligned}
$$

Using Lemma 1 we can estimate $\|r_k\|$ from the last three elements of p_{k+1} and the last two of t_k:

$$
\begin{aligned}
\|r_1\| &= \sqrt{(\bar{\pi}_1^2 - \bar{\tau}_1^2) + \bar{\chi}_2^2}, && (33a)\\
\|r_k\| &= \sqrt{(\bar{\bar{\pi}}_{k-1} - \bar{\bar{\tau}}_{k-1})^2 + (\bar{\pi}_k - \bar{\tau}_k)^2 + \bar{\chi}_{k+1}^2}, \quad k \geq 2. && (33b)
\end{aligned}
$$

Lemma 1. *In (28), $\pi_i = \tau_i$ for $i = 1, \dots, k-2$.*

See proof on page 15.

5.2 Estimating $\|Ar_k\|$

From (21) we have

$$
\|Ar_k\| = \sqrt{\bar{\bar{\zeta}}_{k+1}^2 + \bar{\zeta}_{k+2}^2}, \quad k \leq \ell - 2, \quad \|Ar_{\ell-1}\| = |\zeta_\ell|. \tag{34}
$$

6 cAr

We now introduce cAr, a conjugate direction method in the vein of Cg and Cr of Hestenes and Stiefel [9, 17] for solving $Ax = b$ when A is SPD. By design, cAr is equivalent to MinAres in exact arithmetic as both methods minimize the same quantities in the same subspace, and generate the same iterates. The name cAr stems from the property that successive A-residuals are conjugate with respect to A. The three methods generate sequences of approximate solutions x_k in the Krylov subspaces $\mathcal{K}_k(A, b)$ by minimizing a quadratic function $f(x)$:

$$
\begin{aligned}
f_{\text{Cg}}(x) &= \tfrac{1}{2}x^TAx - b^Tx, & \nabla f_{\text{Cg}}(x) &= -r, & \nabla^2 f_{\text{Cg}}(x) &= A,\\
f_{\text{Cr}}(x) &= \tfrac{1}{2}\|Ax - b\|^2, & \nabla f_{\text{Cr}}(x) &= -Ar, & \nabla^2 f_{\text{Cr}}(x) &= A^2,\\
f_{\text{cAr}}(x) &= \tfrac{1}{2}\|A^2x - Ab\|^2, & \nabla f_{\text{cAr}}(x) &= -A^3r, & \nabla^2 f_{\text{cAr}}(x) &= A^4.
\end{aligned}
$$

Note that all three quadratic functions satisfy $A\nabla f(x) = -\nabla^2 f(x)r$, where $r = b - Ax$. Because cAr minimizes $\|Ar_k\|$ in $\mathcal{K}_k(A, b)$, it is an alternative version of MinAres restricted to SPD A. We can derive it as a descent method with exact linesearch. From initial vectors $x_0 = 0$ and $r_0 = p_0 = b$, we update the iterates with $x_{k+1} = x_k + \alpha_k p_k$. From the Taylor expansion, we can determine α_k that minimizes $f(x_k + \alpha p_k)$:

$$
f(x_k + \alpha p_k) = f(x_k) + \alpha\nabla f(x_k)^T p_k + \tfrac{1}{2}\alpha^2 p_k^T\nabla^2 f(x_k)p_k, \quad \alpha_k = -\frac{\nabla f(x_k)^T p_k}{p_k^T\nabla^2 f(x_k)p_k}.
$$

Afterwards we update the residuals with $r_{k+1} = r_k - \alpha_k Ap_k$ and the directions with $p_{k+1} = r_{k+1} - \sum_{j=0}^{k}\gamma_{k+1,j}p_j$ such that $\text{Span}\{p_0, \dots, p_{k+1}\}$ forms a basis of $\mathcal{K}_{k+2}(A, b)$. We could apply a Gram–Schmidt process to orthogonalize p_{k+1} against all previous directions, but a more relevant approach is to H-conjugate them to derive a shorter recurrence, where $H = \nabla^2 f(x)$ is constant. H-conjugacy also ensures that the vectors are linearly independent. For $i = 0, \dots, k$, $p_i^T H p_{k+1} = 0$ implies $\gamma_{k+1,i} = p_i^T H r_{k+1} / p_i^T H p_i$. Let $\mathcal{P}_k := \mathrm{Span}\{p_0, \dots, p_k\} = \mathrm{Span}\{r_0, \dots, r_k\}$. The exact line-search property yields $\nabla f(x_{k+1})^T p_k$ but also $\nabla f(x_{k+1}) \perp \mathcal{P}_k$ — see, e.g., [14, proof of Theorem 5.2]. Because $Ap_i = (r_i - r_{i+1})/\alpha_i \in \mathrm{Span}\{r_i, r_{i+1}\} \subset \mathcal{P}_k$ for $i = 0, \dots, k-1$, we have $p_i^T A \nabla f(x_{k+1}) = -p_i^T \nabla^2 f(x_{k+1}) r_{k+1} = -p_i^T H r_{k+1} = 0$ and $\gamma_{k+1,i} = 0$. With $\beta_k = -\gamma_{k+1,k} = -p_k^T H r_{k+1} / p_k^T H p_k$, we obtain $p_{k+1} = r_{k+1} + \beta_k p_k$.

Theorem 6. *For* Cg, Cr *and* cAr, *we have:*

$$\alpha_k = \frac{\rho_k}{p_k^T H p_k} \quad \textit{and} \quad \beta_k = \frac{\rho_{k+1}}{\rho_k} \quad \textit{with} \quad \rho_k = -\nabla f(x_k)^T r_k.$$

See proof on page 15.

Cg, Cr and cAr require A to be SPD because we then have $\alpha_k > 0$ until $r_k = 0$. The formulations of Cg (Algorithm 3), Cr (Algorithm 4) and cAr (Algorithm 5) compare the methods and suggest efficient implementations. The vectors $s_k = Ar_k$, $q_k = Ap_k$, $t_k = As_k = A^2 r_k$ and $u_k = Aq_k = A^2 p_k$ ultimately involve just one matrix-vector product with A per iteration. Properties of cAr are summarized in Theorem 7. By virtue of its equivalence to MinAres in exact arithmetic, cAr allows us to establish monotonicity of relevant quantities for MinAres (Theorem 8) on SPD systems. The proofs are strongly inspired by those in [5, 12] for similar properties of Cr and Minres.

Algorithm 3 **CG**

Require: A, b, $\epsilon > 0$
- $k = 0$, $x_0 = 0$
- $r_0 = b$, $p_0 = r_0$
- $q_0 = Ap_0$
- $\rho_0 = r_0^T r_0$
- **while** $\|r_k\| > \epsilon$ **do**
  - $\alpha_k = \rho_k / p_k^T q_k$
  - $x_{k+1} = x_k + \alpha_k p_k$
  - $r_{k+1} = r_k - \alpha_k q_k$
  - $\rho_{k+1} = r_{k+1}^T r_{k+1}$
  - $\beta_k = \rho_{k+1} / \rho_k$
  - $p_{k+1} = r_{k+1} + \beta_k p_k$
  - $q_{k+1} = Ap_{k+1}$
  - $k \leftarrow k + 1$
- **end while**

Algorithm 4 **CR**

Require: A, b, $\epsilon > 0$
- $k = 0$, $x_0 = 0$
- $r_0 = b$, $p_0 = r_0$
- $s_0 = Ar_0$, $q_0 = s_0$
- $\rho_0 = r_0^T s_0$
- **while** $\|r_k\| > \epsilon$ **do**
  - $\alpha_k = \rho_k / \|q_k\|^2$
  - $x_{k+1} = x_k + \alpha_k p_k$
  - $r_{k+1} = r_k - \alpha_k q_k$
  - $s_{k+1} = Ar_{k+1}$
  - $\rho_{k+1} = r_{k+1}^T s_{k+1}$
  - $\beta_k = \rho_{k+1} / \rho_k$
  - $p_{k+1} = r_{k+1} + \beta_k p_k$
  - $q_{k+1} = s_{k+1} + \beta_k q_k$
  - $k \leftarrow k + 1$
- **end while**

Algorithm 5 **CAR**

Require: A, b, $\epsilon > 0$
- $k = 0$, $x_0 = 0$
- $r_0 = b$, $p_0 = r_0$
- $s_0 = Ar_0$, $q_0 = s_0$
- $t_0 = As_0$, $u_0 = t_0$
- $\rho_0 = s_0^T t_0$
- **while** $\|r_k\| > \epsilon$ **do**
  - $\alpha_k = \rho_k / \|u_k\|^2$
  - $x_{k+1} = x_k + \alpha_k p_k$
  - $r_{k+1} = r_k - \alpha_k q_k$
  - $s_{k+1} = s_k - \alpha_k u_k$
  - $t_{k+1} = As_{k+1}$
  - $\rho_{k+1} = s_{k+1}^T t_{k+1}$
  - $\beta_k = \rho_{k+1} / \rho_k$
  - $p_{k+1} = r_{k+1} + \beta_k p_k$
  - $q_{k+1} = s_{k+1} + \beta_k q_k$
  - $u_{k+1} = t_{k+1} + \beta_k u_k$
  - $k \leftarrow k + 1$
- **end while**

Lemma 2. *Let A be SPD. The following properties hold for* cAr *and* MinAres *for all $k \geq 0$:*

(a) $\zeta_{k+1} d_{k+1} = \alpha_k p_k$

(b) $s_k = Ar_k$

(c) $q_k = Ap_k$

(d) $t_k = As_k$

(e) $u_k = Aq_k$.

See proof on page 16.

Theorem 7. *Let A be SPD. For $(i, j) \in \{0, \dots, n-1\}^2$, the following properties hold for* cAr:

(a) $p_i^T A^4 p_j = 0$ *($i \neq j$)*

(b) $r_i^T A^3 p_j = 0$ *($i > j$)*

(c) $r_i^T A^3 r_j = 0$ *(i* $\neq$ *j)*
(d) $\alpha_i \geq 0$
(e) $\beta_i \geq 0$
(f) $q_i^T u_j = p_i^T A^3 p_j \geq 0$
(g) $q_i^T q_j = p_i^T A^2 p_j \geq 0$
(h) $q_i^T p_j = p_i^T A p_j \geq 0$
(i) $p_i^T p_j \geq 0$
(j) $x_i^T p_j \geq 0$
(k) $r_i^T q_j = r_i^T A p_j \geq 0.$

See proof on page 16.

Theorem 8. *For* cAr *(and hence* MinAres*) applied to* $Ax = b$ *when* A *is SPD, the following properties are satisfied:*

- $\|x_k\|$ *increases monotonically*
- $\|x^\star - x_k\|$ *decreases monotonically*
- $\|x^\star - x_k\|_A$ *decreases monotonically*
- $\|r_k\|$ *decreases monotonically.*

See proof on page 17.

7 Implementation and numerical experiments

We implemented Algorithm 2 and Algorithm 5 in Julia [1], version 1.9, as part of our `Krylov.jl` collection of Krylov methods [13]. These implementations of MinAres and cAr are applicable in any floating-point system supported by Julia, including complex numbers, and they run on CPU and GPU. They also support preconditioners.

We evaluate the performance of MinAres on systems generated from symmetric matrices A in the SuiteSparse Matrix Collection [3]. In each case we first scale A to be A/α with $\alpha = \max |A_{ij}|$, so that $\|A\| \approx 1$.

In our first set of experiments, we compare MinAres to our Julia implementation of Minres-qlp in terms of number of iterations on consistent systems when the stopping criterion is $\|r_k\| \leq 10^{-10}$, then when it is $\|Ar_k\| \leq 10^{-10}$. The right-hand side $b = Ae$ (with e a vector of ones) ensures that the system is consistent even if A is singular. The residual and A-residual are calculated explicitly at each iteration in order to evaluate $\|r_k\|$ and $\|Ar_k\|$. (To get a fair comparison, (33) and (34) are not used.) Figure 1 reports residual and A-residual histories for MinAres and Minres-qlp on problems *rail_5177* and *bcsstm36*. We observe that Minres-qlp's $\|Ar_k\|$ is erratic, whereas MinAres's $\|Ar_k\|$ and $\|r_k\|$ are both smooth. Also, Minres-qlp's $\|Ar_k\|$ lags further behind MinAres's than MinAres's $\|r_k\|$ does behind Minres-qlp's. When the system is consistent, we have similar behavior whether A is singular or not.

In a second set of experiments, we compare MinAres to our Julia implementations of Minres-qlp and Lsmr in terms of number of matrix-vector products Av on singular inconsistent systems with $b = e$ when the stopping criterion is $\|Ar_k\| \leq 10^{-6}$ for the problem *zenios* and $\|Ar_k\| \leq 10^{-10}$ for *laser*. Figure 2 shows that Minres-qlp has difficulty reaching the specified $\|Ar_k\|$, but MinAres performs well and converges much faster than Lsmr, the only other Krylov method that minimizes $\|Ar_k\|$.

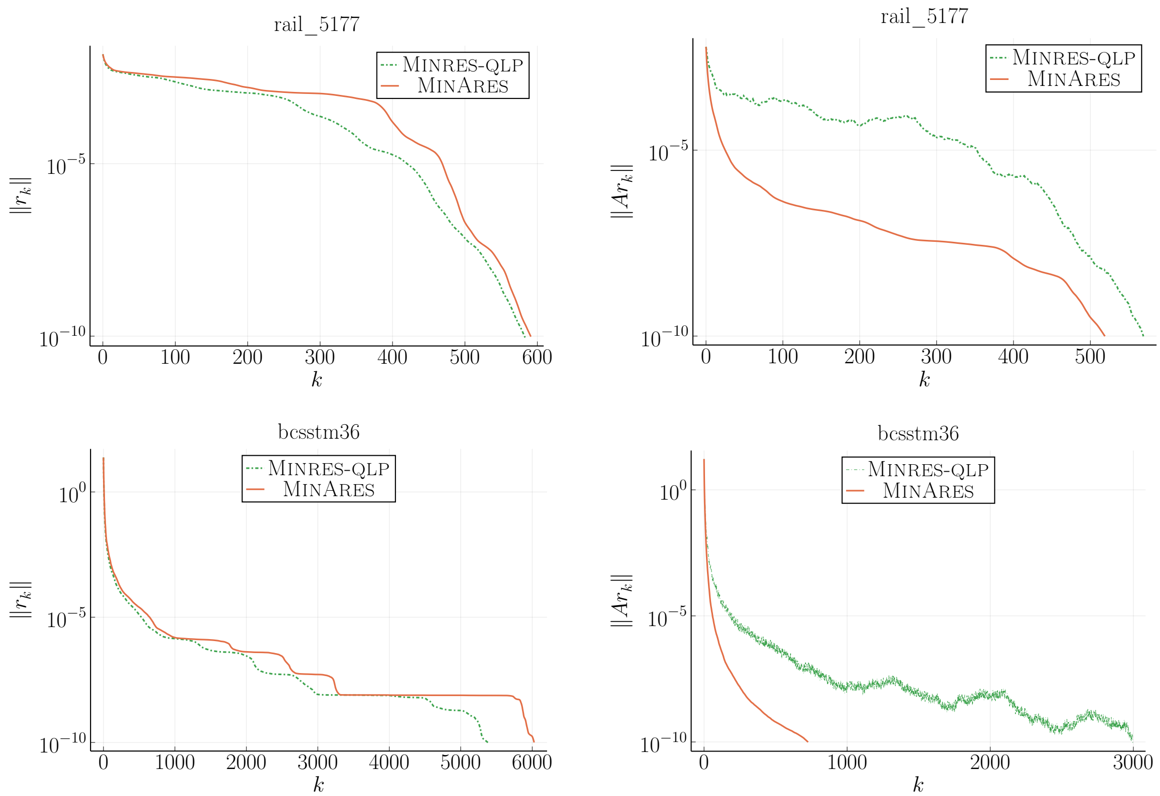

Figure 1: Residual and A-residual histories for MinAres and Minres-qlp on consistent systems generated from the SuiteSparse Matrix Collection. Top: System based on the nonsingular matrix rail_5177 ($n = 5177$). Bottom: System based on the singular matrix bcsstm36 ($n = 23052$)

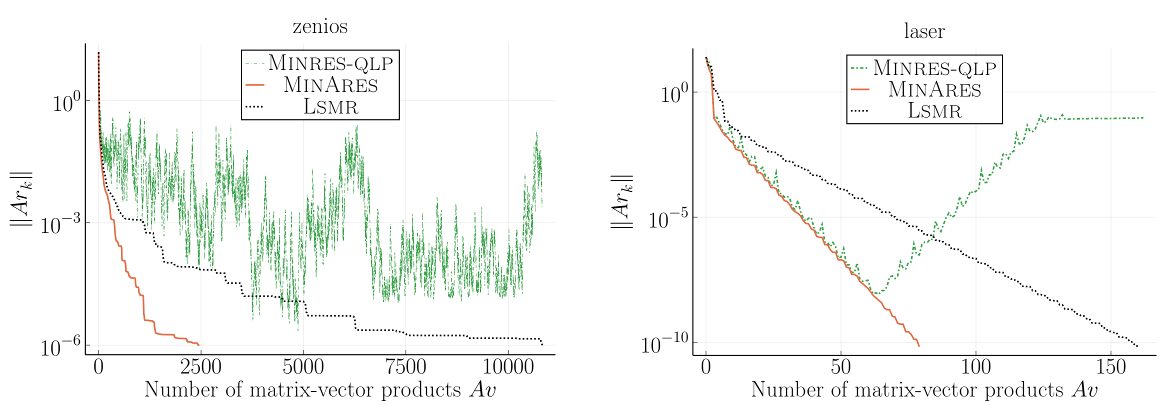

Figure 2: A-residual history for MinAres, Minres-qlp and Lsmr on singular inconsistent systems generated from the SuiteSparse Matrix Collection. Left: System based on the singular matrix zenios ($n = 2873$). Right: System based on the singular matrix laser ($n = 3002$)

8 Summary

MINARES completes the family of Krylov methods based on the symmetric Lanczos process. By minimizing $\|Ar_k\|$ (which always converges to zero), MINARES can be applied safely to any symmetric system. For SPD systems, CAR is equivalent to MINARES and extends the conjugate directions family CG and CR. For such systems we prove that $\|r_k\|$, $\|x_k - x^\star\|$ and $\|x_k - x^\star\|_A$ decrease monotonically for CAR and hence MINARES.

On consistent symmetric systems, MinAres is a relevant alternative to Minres and Minres-qlp because it converges in a similar number of iterations if the stopping condition is based on $\|r_k\|$, and much faster if the stopping condition is based on $\|Ar_k\|$. On singular inconsistent symmetric systems, MinAres outperforms Minres-qlp and Lsmr, and should be the preferred method.

A Proofs

Theorem 1. *For $k \le \ell - 2$, $T_{k+2,k+1}T_{k+1,k}$ has rank k.*

Proof of Theorem 1. From (13a) and (15) we have

$$T_{k+2,k+1}T_{k+1,k} = \begin{bmatrix} R_k^T R_k \\ (\varepsilon_{k-1} e_{k-1}^T + \gamma_k e_k^T) R_k \\ \varepsilon_k e_k^T R_k \end{bmatrix},$$

where $R_k^T R_k$ has rank k because $T_{k+1,k}$ and hence R_k have full column rank. □

Theorem 2. *$T_\ell T_{\ell,\ell-1}$ has rank $\ell - 1$.*

Proof of Theorem 2. From (13b) and (15) we have

$$T_\ell T_{\ell,\ell-1} = \begin{bmatrix} R_{\ell-1}^T R_{\ell-1} \\ (\varepsilon_{\ell-1} e_{\ell-1}^T + \gamma_\ell e_\ell^T) R_{\ell-1} \end{bmatrix},$$

where $R_{\ell-1}^T R_{\ell-1}$ has rank $\ell - 1$ because $T_{\ell,\ell-1}$ and $R_{\ell-1}$ have full column rank. □

Theorem 3. *$T_\ell^2 y_\ell = \beta_1\alpha_1 e_1 + \beta_1\beta_2 e_2$ is consistent even if T_ℓ is singular.*

Proof of Theorem 3. If T_ℓ is singular, the symmetry of T_ℓ and its complete orthogonal decomposition give

$$T_\ell = Q \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix} P = P^T \begin{bmatrix} L^T & 0 \\ 0 & 0 \end{bmatrix} Q^T \quad \text{and} \quad T_\ell^2 = P^T \begin{bmatrix} L^T L & 0 \\ 0 & 0 \end{bmatrix} P,$$

where Q and P are orthogonal and $\operatorname{rank}(L) = \ell - 1$. Thus,

$$\begin{aligned} T_\ell^2 y_\ell - \beta_1\alpha_1 e_1 - \beta_1\beta_2 e_2 &= T_\ell^2 y_\ell - \beta_1 T_\ell e_1 \\ &= P^T \left(\begin{bmatrix} L^T L & 0 \\ 0 & 0 \end{bmatrix} P y_\ell - \beta_1 \begin{bmatrix} L^T & 0 \\ 0 & 0 \end{bmatrix} Q^T e_1 \right) \\ &= P^T \begin{bmatrix} L^T L t_{\ell-1} - L^T u_{\ell-1} \\ 0 \end{bmatrix}, \end{aligned}$$

where $t_{\ell-1}$ and $u_{\ell-1}$ are the first $\ell - 1$ components of $P y_\ell$ and $\beta_1 Q^T e_1$. Because L has full rank, $L^T L t_{\ell-1} = L^T u_{\ell-1}$ has a unique solution. Then, $y_\ell = P^T \left[\begin{smallmatrix} t_{\ell-1} \\ \omega \end{smallmatrix}\right]$ is a solution of $T_\ell^2 y_\ell = \beta_1\alpha_1 e_1 + \beta_1\beta_2 e_2$ for any ω, which means the system is consistent. □

Theorem 4. *If $b \in \operatorname{range}(A)$, the final MinAres iterate x_ℓ is the minimum-length solution of $Ax = b$ (and $r_\ell = b - Ax_\ell = 0$).*

Proof of Theorem 4. The final MinAres subproblem is $T_\ell^2 y_\ell = \beta_1\alpha_1 e_1 + \beta_1\beta_2 e_2 = T_\ell \beta_1 e_1$. Because $b \in \operatorname{range}(A)$, T_ℓ is nonsingular, and the latter system is equivalent to $T_\ell y_\ell = \beta_1 e_1$, the subproblem solved by Minres and Minres-qlp. The final iterate generated by these methods is the minimum-length solution of $Ax = b$ [2, sec. 3.2 theorem 3.1]. □

Theorem 5. *If $Ax = b$ is inconsistent, $\zeta_\ell = 0$ and $Ar_{\ell-1} = 0$.*

Proof of Theorem 5. From (13c), (16) and Theorem 3:

$$z_\ell = \widetilde{Q}_\ell^T(\beta_1\alpha_1 e_1 + \beta_1\beta_2 e_2) = \widetilde{Q}_\ell^T T_\ell^2 y_\ell = \widetilde{Q}_\ell^T N_\ell R_\ell y_\ell = U_\ell R_\ell y_\ell.$$

When $Ax = b$ is inconsistent, T_ℓ has rank $\ell - 1$ and $r_{\ell\ell} = 0$. Because R_ℓ and U_ℓ are upper triangular matrices, $\zeta_\ell = u_{\ell\ell} r_{\ell\ell} \upsilon_\ell = 0$, where υ_ℓ is the last component of y_ℓ. From (21), $Ar_{\ell-1} = 0$ when $\zeta_\ell = 0$. □

Lemma 1. *In (28), $\pi_i = \tau_i$ for $i = 1, \ldots, k-2$.*

Proof of Lemma 1. Let L_{k-2} be the leading $(k-2)\times(k-2)$ submatrix of $\hat{L}_k$, and $J_{m,n}$ be the first m rows of I_n. Then

$$\begin{aligned}
L_{k-2}J_{k-2,k+1}p_{k+1} &= J_{k-2,k}\hat{L}_k J_{k,k+1}\begin{bmatrix}\hat{P}_k & 0\\ 0 & 1\end{bmatrix}Q_k^T\beta_1 e_1\\
&= J_{k-2,k}U_k J_{k,k+1}Q_k^T\beta_1 e_1\\
&= J_{k-2,k+2}\widetilde{Q}_k^T N_k J_{k,k+1}Q_k^T\beta_1 e_1\\
&= J_{k-2,k+2}\widetilde{Q}_k^T T_{k+2,k+1}Q_k J_{k,k+1}^T J_{k,k+1}Q_k^T\beta_1 e_1\\
&= J_{k-2,k+2}\widetilde{Q}_k^T T_{k+2,k+1}Q_k(I_{k+1} - e_{k+1}e_{k+1}^T)Q_k^T\beta_1 e_1\\
&= J_{k-2,k+2}\widetilde{Q}_k^T(\beta_1\alpha_1 e_1 + \beta_1\beta_2 - \bar{\chi}_{k+1}T_{k+2,k+1}Q_k e_{k+1})\\
&= J_{k-2,k+2}(\bar{z}_k - \bar{\chi}_{k+1}\widetilde{Q}_k^T T_{k+2,k+1}Q_k e_{k+1}).
\end{aligned}$$

We now have $T_{k+2,k+1}Q_k e_{k+1} = -(\alpha_{k+1}c_k + \beta_{k+1}c_{k-1}s_k)e_{k+1} - c_k\beta_{k+2}e_{k+2}$. Further, from the structure of the reflections composing $\widetilde{Q}_k^T$, the first $k-2$ elements of $\widetilde{Q}_k^T T_{k+2,k+1}Q_k e_{k+1}$ are zero. Thus,

$$L_{k-2}(\pi_1, \ldots, \pi_{k-2}) = z_{k-2}.$$

Because L_{k-2} is always nonsingular,

$$L_{k-2}\begin{bmatrix}\pi_1 - \tau_1\\ \vdots\\ \pi_{k-2} - \tau_{k-2}\end{bmatrix} = 0 \quad\Longrightarrow\quad \begin{bmatrix}\pi_1\\ \vdots\\ \pi_{k-2}\end{bmatrix} = \begin{bmatrix}\tau_1\\ \vdots\\ \tau_{k-2}\end{bmatrix}.$$

□

Theorem 6. *For* CG, CR *and* cAR, *we have:*

$$\alpha_k = \frac{\rho_k}{p_k^T H p_k} \quad \textit{and} \quad \beta_k = \frac{\rho_{k+1}}{\rho_k} \quad \textit{with} \quad \rho_k = -\nabla f(x_k)^T r_k.$$

Proof of Theorem 6. Let $\rho_k = -\nabla f(x_k)^T r_k$. Because $p_k = r_k + \beta_{k-1}p_{k-1}$ and $\nabla f(x_k) \perp p_{k-1}$ (exact linesearch property), $\nabla f(x_k)^T p_k = \nabla f(x_k)^T r_k$. Therefore,

$$\alpha_k = -\frac{\nabla f(x_k)^T p_k}{p_k^T H p_k} = -\frac{\nabla f(x_k)^T r_k}{p_k^T H p_k} = \frac{\rho_k}{p_k^T H p_k}.$$

Because the directions p_i are H-conjugate, $p_k^T H p_k = p_k^T H(r_k + \beta_{k-1}p_{k-1}) = p_k^T H r_k$. With the relations $Hr_i = -A\nabla f(x_i)$ and $Ap_k = (r_k - r_{k+1})/\alpha_k$, we have:

$$\beta_k = -\frac{p_k^T H r_{k+1}}{p_k^T H p_k} = -\frac{p_k^T H r_{k+1}}{p_k^T H r_k} = -\frac{\nabla f(x_{k+1})^T(r_k - r_{k+1})}{\nabla f(x_k)^T(r_k - r_{k+1})} = \frac{\nabla f(x_{k+1})^T r_{k+1}}{\nabla f(x_k)^T r_k} = \frac{\rho_{k+1}}{\rho_k},$$

where we used the fact that $\nabla f(x_{k+1})^T r_k = -r_{k+1}^T A^i r_k = \nabla f(x_k)^T r_{k+1} = 0$, ($i = 0$ for CG, $i = 1$ for CR and $i = 3$ for cAR). □

Lemma 2. *Let A be SPD. The following properties hold for* cAr *and* MinAres *for all $k \geq 0$:*

(a) $\zeta_{k+1} d_{k+1} = \alpha_k p_k$
(b) $s_k = A r_k$
(c) $q_k = A p_k$
(d) $t_k = A s_k$
(e) $u_k = A q_k$.

Proof of Lemma 2. (a) follows by direct comparison of Algorithm 2 and Algorithm 5.

(b)–(e) all hold by construction at $k = 0$. By induction, assume that they also hold at index $k \geq 0$. Then, $s_{k+1} = s_k - \alpha_k u_k = A r_k - \alpha_k A q_k = A r_{k+1}$, which establishes (b). The remaining properties follow similarly. □

Theorem 7. *Let A be SPD. For $(i, j) \in \{0, \dots, n-1\}^2$, the following properties hold for* cAr*:*

(a) $p_i^T A^4 p_j = 0$ *($i \neq j$)*
(b) $r_i^T A^3 p_j = 0$ *($i > j$)*
(c) $r_i^T A^3 r_j = 0$ *($i \neq j$)*
(d) $\alpha_i \geq 0$
(e) $\beta_i \geq 0$
(f) $q_i^T u_j = p_i^T A^3 p_j \geq 0$
(g) $q_i^T q_j = p_i^T A^2 p_j \geq 0$
(h) $q_i^T p_j = p_i^T A p_j \geq 0$
(i) $p_i^T p_j \geq 0$
(j) $x_i^T p_j \geq 0$
(k) $r_i^T q_j = r_i^T A p_j \geq 0$.

Proof of Theorem 7. Because $\nabla^2 f_{\text{cAr}}(x) = A^4$, we A^4-conjugate the vectors p_i by construction and (a) is satisfied.

Because $\nabla f_{\text{cAr}}(x_i) = -A^3 r_i$, the exact linesearch property yields (b) as in [14, proof of Theorem 5.2].

If $i > j$, $r_i^T A^3 r_j = r_i^T A^3 (p_j - \beta_{j-1} p_{j-1}) = 0$ by (b). If $i < j$, $r_i^T A^3 r_j = (p_i - \beta_{i-1} p_{i-1})^T A^3 r_j = 0$, again thanks to (b), which proves (c).

First note that $\rho_i = s_i^T t_i = r_i^T A^3 r_i \geq 0$ because A is SPD. Thus $\alpha_i = \rho_i / \|u_i\|^2 \geq 0$ and $\beta_i = \rho_{i+1}/\rho_i \geq 0$, which proves (d) and (e).

We now establish (f) by induction. If $i = j$, $q_i^T u_i = q_i^T A q_i \geq 0$ because A is SPD. Assuming $q_i^T u_j \geq 0$ when $|i - j| = k - 1 \geq 0$, we want to show the result for $|i - j| = k$. If $i - j = k > 0$ then $q_i^T u_j = q_i^T u_{i-k}$. Otherwise we have $j - i = k > 0$ and $q_i^T u_j = q_i^T u_{i+k}$. Lemma 2 yields

$$\begin{aligned} q_i^T u_{i-k} &= (s_i + \beta_{i-1} q_{i-1})^T u_{i-k} \\ &= s_i^T u_{i-k} + \beta_{i-1} q_{i-1}^T u_{i-k} \\ &= r_i^T A^3 p_{i-k} + \beta_{i-1} q_{i-1}^T u_{i-k} \\ &= \beta_{i-1} q_{i-1}^T u_{i-k} \end{aligned} \qquad \begin{aligned} q_i^T u_{i+k} &= q_i^T A q_{i+k} \\ &= q_i^T A (s_{i+k} + \beta_{i+k-1} q_{i+k-1}) \\ &= p_i^T A^3 r_{i+k} + \beta_{i+k-1} u_i^T q_{i+k-1} \\ &= \beta_{i+k-1} q_{i+k-1}^T u_i \end{aligned}$$

$\beta_{i-1} \geq 0$ and $\beta_{i+k-1} \geq 0$ by (e). $q_{i-1}^T u_{i-k} \geq 0$ and $q_{i+k-1}^T u_i \geq 0$ by induction assumption. Thus, $q_i^T u_j \geq 0$ for $|i-j| = k$, which completes the proof of (f).

At termination, define $\mathcal{P} = \text{Span}\{p_0, \dots, p_{\ell-1}\}$, $\mathcal{Q} = \text{Span}\{q_0, \dots, q_{\ell-1}\} = A\mathcal{P}$ and $\mathcal{U} = \text{Span}\{u_0, \dots, u_{\ell-1}\} = A\mathcal{Q}$. By construction, $\mathcal{P} = \text{Span}\{b, \dots, A^{\ell-1}b\}$, $\mathcal{Q} = \text{Span}\{Ab, \dots, A^{\ell}b\}$ and $\mathcal{U} = \text{Span}\{A^2 b, \dots, A^{\ell+1}b\}$. Again by construction, $x_\ell \in \mathcal{P}$, and since $r_\ell = 0$, we have $Ax_\ell = b \in \mathcal{Q}$ and $A^2 x_\ell = Ab \in \mathcal{U}$. We see that $\mathcal{P} \subset \mathcal{Q} \subset \mathcal{U}$.

(a) and Lemma 2 (c)–(e) imply that $u_i^T u_j = 0$ for $i \neq j$, and therefore, $\{u_k/\|u_k\|\}_{k=0,\dots,\ell-1}$ forms an orthonormal basis for $\mathcal{U}$. Thus, if we project p_i and q_i into $\mathcal{U}$, we have

$$p_i = \sum_{k=0}^{\ell-1} \frac{p_i^T u_k}{u_k^T u_k} u_k \quad \text{and} \quad q_i = \sum_{k=0}^{\ell-1} \frac{q_i^T u_k}{u_k^T u_k} u_k.$$

Scalar products between these vectors can be expressed as

$$q_i^T q_j = \sum_{k=0}^{\ell-1} \frac{(q_i^T u_k)(q_j^T u_k)}{\|u_k\|^2}, \quad p_i^T q_j = \sum_{k=0}^{\ell-1} \frac{(p_i^T u_k)(q_j^T u_k)}{\|u_k\|^2} \quad \text{and} \quad p_i^T p_j = \sum_{k=0}^{\ell-1} \frac{(p_i^T u_k)(p_j^T u_k)}{\|u_k\|^2}.$$

Thus $q_i^T q_j \geq 0$ by (f), proving (g). Because $p_i^T u_k = p_i^T A q_k = q_i^T q_k$, $p_i^T q_j \geq 0$ and $p_i^T p_j \geq 0$ by (f) and (g), which proves (h) and (i).

By construction, $x_i = \sum_{k=0}^{i} \alpha_k p_k$ and so $x_i^T p_j \geq 0$ by (d) and (i), proving (j).

Finally, $r_i^T q_j = \sum_{k=i}^{\ell-1} \alpha_k q_k^T q_j \geq 0$ by (d) and (g), proving (k). □

Theorem 8. *For* CAR *(and hence* MINARES*) applied to $Ax = b$ when A is SPD, the following properties are satisfied:*

- $\|x_k\|$ *increases monotonically*
- $\|x^\star - x_k\|$ *decreases monotonically*
- $\|x^\star - x_k\|_A$ *decreases monotonically*
- $\|r_k\|$ *decreases monotonically.*

Proof of Theorem 8. From Theorem 7 (d) and (j),

$$\begin{aligned}\|x_k\|^2 - \|x_{k-1}\|^2 &= (x_{k-1} + \alpha_k p_k)^T (x_{k-1} + \alpha_k p_k) - x_{k-1}^T x_{k-1} \\ &= 2\alpha_k p_k^T x_{k-1} + \alpha_k^2 \|p_k\|^2 \geq 0.\end{aligned}$$

From Theorem 7 (d) and (i),

$$\begin{aligned}\|x^\star - x_{k-1}\|^2 - \|x^\star - x_k\|^2 &= \left(\sum_{i=k}^{\ell-1} \alpha_i p_i\right)^T \left(\sum_{i=k}^{\ell-1} \alpha_i p_i\right) - \left(\sum_{i=k+1}^{\ell-1} \alpha_i p_i\right)^T \left(\sum_{i=k+1}^{\ell-1} \alpha_i p_i\right) \\ &= 2\alpha_k p_k^T \left(\sum_{i=k+1}^{\ell-1} \alpha_i p_i\right) + \alpha_k^2 \|p_k\|^2 \geq 0.\end{aligned}$$

From Theorem 7 (d) and (h),

$$\begin{aligned}\|x^\star - x_{k-1}\|_A^2 - \|x^\star - x_k\|_A^2 &= \left(\sum_{i=k}^{\ell-1} \alpha_i p_i\right)^T A \left(\sum_{i=k}^{\ell-1} \alpha_i p_i\right) - \left(\sum_{i=k+1}^{\ell-1} \alpha_i p_i\right)^T A \left(\sum_{i=k+1}^{\ell-1} \alpha_i p_i\right) \\ &= 2\alpha_k q_k^T \left(\sum_{i=k+1}^{\ell-1} \alpha_i p_i\right) + \alpha_k^2 q_k^T p_k \geq 0.\end{aligned}$$

From Theorem 7 (d) and (k),

$$\begin{aligned}\|r_{k-1}\|^2 - \|r_k\|^2 &= r_{k-1}^T r_{k-1} - r_k^T r_k \\ &= (r_k + \alpha_{k-1} q_{k-1})^T (r_k + \alpha_{k-1} q_{k-1}) - r_k^T r_k \\ &= 2\alpha_{k-1} q_{k-1}^T r_k + \alpha_{k-1}^2 \|q_{k-1}\|^2 \geq 0. \end{aligned}$$

□

References

[1] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. SIAM Rev., 59(1):65–98, 2017.

[2] S.-C. Choi, C. C. Paige, and M. A. Saunders. MINRES-QLP: A Krylov subspace method for indefinite or singular symmetric systems. SIAM J. Sci. Comput., 33(4):1810–1836, 2011.

[3] T. Davis and Y. Hu. The University of Florida sparse matrix collection. ACM Trans. Math. Software, 38 (1):1–25, 2011. See `https://sparse.tamu.edu/`.

[4] D. C.-L. Fong and M. A. Saunders. LSMR: An iterative algorithm for sparse least-squares problems. SIAM J. Sci. Comput., 33(5):2950–2971, 2011.

[5] D. C.-L. Fong and M. A. Saunders. CG versus MINRES: an empirical comparison. Sultan Qaboos University Journal for Science, 17(1):44–62, 2012.

[6] K. A. Gallivan, S. Thirumalai, P. V. Dooren, and V. Vermaut. High performance algorithms for Toeplitz and block Toeplitz matrices. Linear Algebra and its Applications, 241:343–388, 1996.

[7] G. H. Golub and W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. SIAM J. Numer. Anal., 2(2):205–224, 1965.

[8] G. H. Golub and C. F. Van Loan. Matrix Computations. The Johns Hopkins University Press, fourth edition, 2013.

[9] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. J. Res. Natl. Bur. Stand., 49(6):409–436, 1952.

[10] E. F. Kaasschieter. Preconditioned conjugate gradients for solving singular systems. J. Computational and Applied Mathematics, 24(1-2):265–275, 1988.

[11] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J. Res. Natl. Bur. Stand., 45:225–280, 1950.

[12] D. G. Luenberger. The conjugate residual method for constrained minimization problems. SIAM J. Numer. Anal., 7(3):390–398, 1970.

[13] A. Montoison and D. Orban. Krylov.jl: A Julia basket of hand-picked Krylov methods. Journal of Open Source Software, 8(89):5187, 2023.

[14] J. Nocedal and S. J. Wright. Numerical Optimization. Springer New York, NY, 2 edition, 2006.

[15] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. SIAM J. Numer. Anal., 12(4):617–629, 1975.

[16] 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.

[17] E. Stiefel. Relaxationsmethoden bester strategie zur lösung linearer gleichungssysteme. Commentarii Mathematici Helvetici, 29(1):157–179, 1955.

[18] L. N. Trefethen and D. Bau III. Numerical Linear Algebra. SIAM, Philadelphia, 1997.