A MODIFIED IMPLEMENTATION OF MINRES TO MONITOR RESIDUAL SUBVECTOR NORMS FOR BLOCK SYSTEMS

Roland Herzog∗ Kirk Soodhalter†

August 2, 2017

Saddle-point systems, i.e., structured linear systems with symmetric matrices are considered. A modified implementation of (preconditioned) MINRES is derived which allows subvectors of the residual to be monitored individually. Compared to the implementation from the textbook of [Elman, Silvester and Wathen, Oxford University Press, 2014], our method requires one extra vector of storage and no additional applications of the preconditioner. Numerical experiments are included.

KEYWORDS: MINRES, saddle-point problems, structured linear systems, preconditioning, subvector norms

1 INTRODUCTION

We are solving symmetric linear systems of the form

\[ \mathbf{K} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} := \begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & -\mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_u \\ \mathbf{f}_p \end{bmatrix}, \]

(1.1)

where \( \mathbf{A} \in \mathbb{R}^{m \times m} \) and \( \mathbf{C} \in \mathbb{R}^{p \times p} \) are symmetric, and \( \mathbf{B} \in \mathbb{R}^{p \times m} \), by applying the preconditioned MINRES [Paige, Saunders, 1975] iteration. After \( j \) iterations, we have an

∗Technische Universität Chemnitz, Faculty of Mathematics, Professorship Numerical Mathematics (Partial Differential Equations), D–09107 Chemnitz, Germany, roland.herzog@mathematik.tu-chemnitz.de, http://www.tu-chemnitz.de/herzog

†Industrial Mathematics Institute, Johannes Kepler University, Altenbergerstraße 69, A–4040 Linz, Austria (kirk.soodhalter@indmath.uni-linz.ac.at, https://www.indmath.uni-linz.ac.at/index.php/members/55-soodhalter).
approximation \((u^{(j)}, p^{(j)})\) and residual

\[
    r^{(j)} = \begin{bmatrix} f_u \\ f_p \end{bmatrix} - \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} u^{(j)} \\ p^{(j)} \end{bmatrix} = \begin{bmatrix} f_u \\ f_p \end{bmatrix} - \begin{bmatrix} Au^{(j)} + B^T p^{(j)} \\ Bu^{(j)} - Cp^{(j)} \end{bmatrix}.
\]

We denote the two parts of the residual as

\[
    r_u^{(j)} = f_u - (Au^{(j)} + B^T p^{(j)}) \quad \text{and} \quad r_p^{(j)} = f_p - (Bu^{(j)} - Cp^{(j)}).
\]

The question we seek to answer is: can we monitor the individual preconditioned norms of \(r_u^{(j)}\) and \(r_p^{(j)}\) (as opposed to the full norm of \(r^{(j)}\)) using only quantities arising in an efficient implementation of the preconditioned MINRES algorithm, namely that in Elman, Silvester, Wathen, 2014, Algorithm 4.1? The answer is yes, under certain conditions. The technical requirements to do so are connected to the notion of so-called “natural” preconditioners (i.e., block diagonal preconditioners) which arise in this setting; see, e.g., Pestana, Wathen, 2015; Zulehner, 2011 and references contained therein. We also note that efficient computation of error and residual norms in this setting have also been treated in the paper of Silvester and Simoncini Silvester, Simoncini, 2011. In this paper, the fact that the error and residual norms are equivalent is exploited, with the authors approximating the lower- and upper-bounding equivalence constants relating the unknown error norm to the computed residual norm. From this, the authors were able to derive a stopping criterion based on a tolerance for the preconditioner-induced error norm. However, this was only for the full residual norm. In Section 2, we demonstrate that at the storage cost of one additional full-length vector and six scalars but no additional applications of the preconditioner, one can modify the preconditioned MINRES method to calculate these norms in a progressive fashion. An extension to more than two residual parts is straightforward. An implementation of our modified version of MINRES is given in Algorithm 1 and is available at Herzog, Soodhalter, 2016 as a Matlab file.

As a motivation for our study, we mention that the individual parts of the residual in (1.1) often have different physical interpretations. Monitoring them individually allows a better insight into the convergence of MINRES and it allows the formulation of refined stopping criteria. We present examples and numerical experiments in Section 4.

## 2 How to Monitor Both Parts of the Preconditioned Residual

In this section, we derive how one monitors these norms without incurring much extra computational or storage expense. We will describe everything in terms of the variable names used in Elman, Silvester, Wathen, 2014, Algorithm 4.1 with two exceptions, which will be noted below.
2.1 The Minimum Residual Method

Here we describe briefly the derivation of preconditioned MINRES as a Krylov subspace method, specifically in order to relate certain quantities from Elman, Silvester, Wathen, 2014, Algorithm 4.1 to the common quantities arising in the Lanczos process. In principle, the preconditioned MINRES algorithm is an implementation of the minimum residual Krylov subspace method when the preconditioned operator is self-adjoint with respect to the inner product induced by the symmetric, positive-definite preconditioner. It can be equivalently formulated as an iteration for operator equations in which the operator is self-adjoint, mapping elements of a Hilbert space into its dual; see Günnel, Herzog, Sachs, 2014. We present the derivation for the finite dimensional case. For the purposes of this discussion, let \( x^{(j)} = (u^{(j)}, p^{(j)}) \) be the \( j \)th approximation.

Let \( K = \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} \). We further assume that the preconditioner \( P \) is symmetric positive definite and has block diagonal structure

\[
P = \begin{bmatrix} P_u & P_p \\ P_u & P_p \end{bmatrix}
\]

with \( P_u \in \mathbb{R}^{m \times m} \) and \( P_p \in \mathbb{R}^{p \times p} \).

As was shown in Günnel, Herzog, Sachs, 2014, the choice of a symmetric positive definite preconditioner \( P \) and the choice of the inner product in the space of iterates are one and the same thing. It is therefore natural to assume the block-diagonal structure of \( P \) when we acknowledge that the residual subvectors often have different physical meaning and need to be measured in different norms; see Pestana, Wathen, 2015; Zulehner, 2011 and Section 4 for examples.

We briefly review some basic facts about Krylov subspaces for symmetric matrices. We describe the situation for right preconditioning but would obtain the same results if left preconditioning were chosen. For some starting element \( h \in \mathbb{R}^n \), we define the \( j \)th Krylov subspace generated by the matrix \( KP^{-1} \) and \( h \) to be

\[
\mathcal{K}_j(KP^{-1}, h) = \text{span}\{h, (KP^{-1})h, (KP^{-1})^2h, \ldots, (KP^{-1})^j h\}.
\]

In the case that \( P \) is a symmetric positive definite preconditioner, it has been shown that a Krylov subspace can be constructed using the short-term Lanczos iteration and that a MINRES method can be used for solving (1.1). In, e.g., Elman, Silvester, Wathen, 2014, preconditioned MINRES is derived by first observing that \( P \) admits the Cholesky decomposition \( P = HH^T \), meaning one can consider solving the two-sided preconditioned equations

\[
H^{-1}KH^{-T}y = H^{-1}b \quad \text{with} \quad y = H^T x \quad \text{where} \quad x = \begin{bmatrix} u \\ u \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} f_u \\ f_p \end{bmatrix},
\]

where we have the initial approximation \( y^{(0)} = H^T x^{(0)} \). Using the Lanczos process, we obtain the matrix \( V_j = [v_1, v_2, \ldots, v_j] \in \mathbb{R}^{n \times j} \) whose orthonormal columns (w.r.t.
the $P^{-1}$ inner product) form a basis of $K_j(KP^{-1}, h)$, and which satisfies the Lanczos relation
\[ KP^{-1}V_j = V_{j+1}\bar{T}_j \quad \text{where} \quad \bar{T}_j \in \mathbb{R}^{(j+1) \times j}. \] (2.1)

The matrix $\bar{T}_j$ is tridiagonal, and the matrix $T_j$ (defined as the first $j$ rows of $\bar{T}_j$) is symmetric, leading to the following well-known three-term recurrence formula
\[ KP^{-1}v_j = \gamma_{j+1}v_{j+1} + \delta_jv_j + \gamma_jv_{j-1}. \] (2.2)

Using the naming conventions in Elman, Silvester, Wathen, 2014, Algorithm 2.4, we have
\[ T_j = \begin{bmatrix} \delta_1 & \gamma_2 & & & \\ \gamma_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ \gamma_j & \ddots & \ddots & \ddots & \gamma_j \\ \end{bmatrix}. \]

The matrix $H^{-1}K^{-1}H^{-T}$ is still symmetric. However, one can notice that the preconditioned residual satisfies $\|H^{-1}r^{(k)}\| = \|r^{(k)}\|_{P^{-1}}$, where $\|\cdot\|_{P^{-1}}$ is the norm arising from the inner product induced by $P^{-1}$, i.e., $\langle \cdot, \cdot \rangle_{P^{-1}} = \langle P^{-1} \cdot, \cdot \rangle$. One further notes that we have the equivalence
\[ K_j(H^{-1}K^{-1}H^{-T}, H^{-1}r^{(0)}) = H^{-1}K_j(KP^{-1}, r^{(0)}). \]

If preconditioned MINRES at iteration $j$ produces a correction
\[ s^{(j)} \in K_j(H^{-1}K^{-1}H^{-T}, H^{-1}r^{(0)}) \quad \text{such that} \quad y^{(j)} = y^{(0)} + s^{(j)}, \]

then we recover the approximation for $x$ with $x^{(j)} = H^{-T}(y^{(0)} + s^{(j)})$. Since $x^{(0)} = H^{-T}y^{(0)}$ the correction space for preconditioned MINRES with respect to the original variables is actually $P^{-1}K_j(KP^{-1}, r^{(0)})$. From here, one can show that the preconditioned MINRES iteration is equivalent to an iteration with the subspace $K_j(KP^{-1}, r^{(0)})$ but with respect to the inner product $\langle \cdot, \cdot \rangle_{P^{-1}}$ induced by the preconditioner. This derivation leads to the implementation of the preconditioned MINRES method shown in, e.g., Elman, Silvester, Wathen, 2014, Algorithm 4.1; see also Günnel, Herzog, Sachs, 2014 for an alternate derivation which avoids the temporary introduction of Cholesky factors.

Let the columns of $V_j$ form an orthonormal basis for $K_j(KP^{-1}, r^{(0)})$ with respect to $\langle \cdot, \cdot \rangle_{P^{-1}}$. Let $Z_j = P^{-1}V_j$ have as columns the image of those vectors under the action of the preconditioner. Then we have the preconditioned Lanczos relation,
\[ KP^{-1}V_j = KZ_j = V_{j+1}\bar{T}_j \] (2.3)
where $T_j$ has the tridiagonal structure described earlier. We have the three-term recurrence

$$Kz_j = \gamma_{j+1}v_{j+1} + \delta_j v_j + \gamma_j v_{j-1}. \quad (2.4)$$

The preconditioned MINRES method solves the following residual minimization problem,

$$x^{(j)} = x^{(0)} + Z_j y^{(j)} \quad \text{where} \quad y^{(j)} = \arg \min_{y \in \mathbb{R}^j} \| \gamma_1 e_1 - T_j y \|_2, \quad (2.5)$$

where $\gamma_1 = \| r^0 \|_{P^{-1}}$.

### 2.2 Implementation of MINRES

Though this reduces the minimization of the residual to a small least-squares problem, this is not the most efficient way to implement the MINRES method; and indeed, this is not what is done in, e.g., Elman, Silvester, Wathen, 2014, Algorithm 4.1. Due to (2.2), MINRES can be derived such that only six full-length vectors must be stored. In the interest of not again deriving everything, we will simply provide a few relationships between quantities from the above explanation of MINRES and those in Elman, Silvester, Wathen, 2014, Algorithm 4.1. Further implementation details can be found in, e.g., Greenbaum, 1997, Section 2.5.

Let $\overline{T}_j = Q_j R_j$ be a QR-factorization of $\overline{T}_j$ computed with Givens rotations where $Q_j \in \mathbb{R}^{(j+1) \times (j+1)}$ is a unitary matrix constructed from the product of Givens rotations, and $R_j \in \mathbb{R}^{(j+1) \times j}$ is upper triangular. The matrix $R_j \in \mathbb{R}^{j \times j}$ is simply $R_j$ with the last row (of zeros) deleted. Since $\overline{T}_j$ is tridiagonal, we must annihilate only one subdiagonal entry per column. Following from Elman, Silvester, Wathen, 2014, Algorithm 4.1, we denote $s_i$ and $c_i$ to be the Givens sine and cosine. We denote

$$Q_j^T = G^{(j)}_{j+1} G^{(j)}_j \ldots G^{(j)}_2$$

with

$$G^{(j)}_i = \begin{bmatrix} I_{i-2} & \begin{bmatrix} c_i & s_i \\ -s_i & c_i \end{bmatrix} & \begin{bmatrix} 1_{j-i} \end{bmatrix} \end{bmatrix} \in \mathbb{R}^{(j+1) \times (j+1)}.$$  

Using a normal equations formulation of (2.5), we can derive an expression for the least squares minimizer

$$y^{(j)} = R_j^{-1} \left\{ Q_j^T (\gamma_1 e_1) \right\}_{1:j'}, \quad (2.6)$$

where $\left\{ \cdot \right\}_{1:j'}$ indicates we take only the first $j$ rows of the argument. For the purpose of discussion, we make an additional modification to the variable naming using in Elman, Silvester, Wathen, 2014, Algorithm 4.1. We index $\eta$, which is used to track the residual norm. On Line 4 of the algorithm, we would have $\eta_0 = \gamma_1$, and at Line 18, we would write $\eta_j = -s_{j+1}\eta_j$. From the proof of Fischer, 1996, Corollary 2.5.3, one can see that at iteration $j$, the least squares residual can be written

$$\gamma_1 e_1 - \overline{T}_j y^{(j)} = Q_j (\eta_j e_{j+1}), \quad (2.7)$$
where \( e_{j+1} \) is the last column of the \( (j+1) \times (j+1) \) identity matrix, and because \( Q_j \) is unitary, it follows that \( \| \eta_j \| \) is the norm of the residual \( \| r^{(j)} \|_{p-1} \) associated to \( x^{(j)} \).

We now derive an expression for the norms of \( r_u^{(j)} \) and \( r_p^{(j)} \) in terms of quantities arising in the Lanczos process and residual minimization. Due to the block structure of \( K \), we partition \( V_j \) and \( Z_j \) similarly,

\[
V_j = \begin{bmatrix} V_{j,u} \\ V_{j,p} \end{bmatrix} \quad \text{and} \quad Z_j = \begin{bmatrix} Z_{j,u} \\ Z_{j,p} \end{bmatrix}
\]  

and then insert the partitioned vectors into (2.3)

\[
\begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} Z_{j,u} \\ Z_{j,p} \end{bmatrix} = \begin{bmatrix} AZ_{j,u} + B^T Z_{j,p} \\ BZ_{j,u} - CZ_{j,p} \end{bmatrix} = \begin{bmatrix} V_{j+1,u} \\ V_{j+1,p} \end{bmatrix} T_j = \begin{bmatrix} V_{j+1,u} \ T_j \\ V_{j+1,p} \ T_j \end{bmatrix}.
\]  

We note that for tracking and updating the full residual vector, this partition is artificially imposed. One could simply track the full residual and partition at the end to compute the subvector norms separately. We derive everything using this block partition because one may wish to track the norm of only one subvector. Furthermore, one needs the partition quantities to define the recursions for updating the preconditioned subvector norms. It then follows that we have Lanczos relations for the blocks

\[
AZ_{j,u} + B^T Z_{j,p} = V_{j+1,u} T_j \quad \text{and} \quad BZ_{j,u} - CZ_{j,p} = V_{j+1,p} T_j.
\]  

From Fischer, 1996, Theorem 2.5.7, we know that the \( j \)th MINRES residual satisfies the recursion

\[
r^{(j)} = r^{(j-1)} - c_{j+1} \eta_{j-1} m^{(j+1)}_j,
\]  

where the auxiliary vectors are columns of the matrix

\[
M_{j+1} = V_{j+1} Q_j = \begin{bmatrix} \ast & \ast & \cdots & \ast \end{bmatrix} \begin{bmatrix} m^{(j+1)}_j \\ m_{j+1} \end{bmatrix}.
\]

The columns of \( M \) can be progressively updated according to

\[
m^{(j)}_j = c_{j+1} m^{(j)}_j + s_{j+1} v_{j+1} \quad \text{and} \quad m^{(j+1)}_{j+1} = -s_{j+1} m^{(j)}_j + c_{j+1} v_{j+1},
\]

where we set \( m^{(1)}_1 = v_1 \). From (2.7), we then have that

\[
r^{(j)} = \eta_j V_{j+1} Q_j e_{j+1} = \eta_j m^{(j+1)}_{j+1}.
\]

If we then partition the vector \( m^{(j+1)}_{j+1} = \begin{bmatrix} m^{(j+1)}_{j+1,u} \\ m^{(j+1)}_{j+1,p} \end{bmatrix} \), we then have formulas for the partial residuals,

\[
r_u^{(j)} = \eta_j m^{(j+1)}_{j+1,u} \quad \text{and} \quad r_p^{(j)} = \eta_j m^{(j+1)}_{j+1,p}.
\]

We can also rewrite (2.11) in order to recover (2.10). From (2.10) it follows that the residual subvectors \( r_u^{(j)} \) and \( r_p^{(j)} \) satisfy the progressive updating formulas

\[
r_u^{(j)} = r_u^{(j-1)} - c_{j+1} \eta_{j-1} m^{(j+1)}_{j,u} \quad \text{and} \quad r_p^{(j)} = r_p^{(j-1)} - c_{j+1} \eta_{j-1} m^{(j+1)}_{j,p}.
\]

However, this representation would require storage of two full-length vectors.
2.3 Minimum residual subvector monitoring

We have seen that through recursion formulas, we can compute one or both subvectors of the residual vector produced by preconditioned MINRES. However, we wish to avoid storing or re-evaluating these vectors since we actually want to compute only the appropriate norm of each piece. The full residual minimization is with respect to the \( \| \cdot \|_{P^{-1}} \). We have assumed in this paper that the preconditioner has block diagonal structure; and as \( P \) is symmetric positive definite, it follows that its blocks are as well. Thus we can write

\[
\| r^{(j)} \|_{P^{-1}}^2 = \| r_u^{(j)} \|_{P_u^{-1}}^2 + \| r_p^{(j)} \|_{P_p^{-1}}^2.
\]

These norms can also be computed recursively using information already on hand. We note that doing it as in the following allows us to avoid additional applications of the preconditioner.

In order to compute the norms separately, it is necessary that we also use the same fact for the newest Lanczos vector,

\[
\| v_{j+1} \|_{P^{-1}}^2 = \| v_{j+1,u} \|_{P_u^{-1}}^2 + \| v_{j+1,p} \|_{P_p^{-1}}^2.
\]

Let

\[
\psi_{j+1,u} = \| v_{j+1,u} \|_{P_u^{-1}}^2 = \langle z_{j+1,u}, v_{j+1,u} \rangle \quad \text{and} \quad \psi_{j+1,p} = \| v_{j+1,p} \|_{P_p^{-1}}^2 = \langle z_{j+1,p}, v_{j+1,p} \rangle.
\]

We can store these values so they are available for later use. We further denote \( \eta_{j,u} = \| r_u^{(j)} \|_{P_u^{-1}} \) and \( \eta_{j,p} = \| r_p^{(j)} \|_{P_p^{-1}} \). This allows us to concisely state the following theorem.

**Theorem 2.1.** The ratio between the full preconditioned residual norm and the respective preconditioned residual subvector norms can be written progressively as follows:

\[
\left( \frac{\eta_{j,u}}{\eta_j} \right)^2 = \frac{2}{s_{j+1}} \left( \frac{\eta_{j-1,u}}{\eta_j} \right)^2 - 2s_{j+1} c_{j+1} \left( \begin{bmatrix} m_{j,u}^{(j)} \end{bmatrix}^T z_{j+1,u} + c_{j+1}^2 \psi_{j+1,u} \right) \quad \text{and}
\]

\[
\left( \frac{\eta_{j,p}}{\eta_j} \right)^2 = \frac{2}{s_{j+1}} \left( \frac{\eta_{j-1,p}}{\eta_j} \right)^2 - 2s_{j+1} c_{j+1} \left( \begin{bmatrix} m_{j,p}^{(j)} \end{bmatrix}^T z_{j+1,p} + c_{j+1}^2 \psi_{j+1,p} \right).
\]

**Proof.** The proofs for each subvector are identical, so we prove it only for one. We can verify the identity by direct calculation and then by showing that the quantities arising
from the calculation are available by recursion. We compute

\[ \|r^{(j)}_u\|_{P_u^{-1}}^2 = \langle \eta_j m^{(j+1)}_{j+1,u}, P_u^{-1}(\eta_j m^{(j)}_{j+1,u}) \rangle \]

\[ = \eta_j^2 \left( -s_{j+1}m^{(j)}_{j,u} + c_{j+1}v_{j+1,u} \right)^T P_u^{-1} \left( -s_{j+1}m^{(j)}_{j,u} + c_{j+1}v_{j+1,u} \right) \]

\[ = \eta_j^2 \left[ s_{j+1}^2 (m^{(j)}_{j,u})^T P_u^{-1} m^{(j)}_{j,u} - 2s_{j+1}c_{j+1} (m^{(j)}_{j,u})^T P_u^{-1} v_{j+1,u} \right. \]

\[ \left. + c_{j+1}^2 v_{j+1,u}^T P_u^{-1} v_{j+1,u} \right] \]

\[ \]

\[ = \eta_j^2 \left[ s_{j+1}^2 \eta_{j-1,u}^2 - 2s_{j+1}c_{j+1} \left( m^{(j)}_{j,u} \right)^T z_{j+1,u} + c_{j+1}^2 \psi_{j+1,u} \right] . \]

Basic algebra finishes the proof.

For completeness, we now present a modified version of Elman, Silvester, Wathen, 2014, Algorithm 4.1, here called Algorithm 1. Note that we omit superscripts for \( m^{(\ell)}_{i,p} \) and some other subscripts where they are not needed. Furthermore, we introduce the scalars \( \theta_u \) and \( \theta_p \) defined by

\[ \theta_u = \langle m^{(j)}_{j,u}, z_{j+1,u} \rangle \quad \text{and} \quad \theta_p = \langle m^{(j)}_{j,p}, z_{j+1,p} \rangle \quad (2.12) \]

and the squared residual norm fractions

\[ \mu_u = \left( \frac{\eta_{j,u}}{\eta_j} \right)^2 \quad \text{and} \quad \mu_p = \left( \frac{\eta_{j,p}}{\eta_j} \right)^2 . \quad (2.13) \]

Notice that one can work with full-length vectors and the partitioning is only important for the computation of partial inner products as in (2.12).

**Remark 2.2.** Note that up to the end of Section 2.2, the assumption that the preconditioner is block diagonal is not used. However, in Section 2.3 it allows us to write the squared full residual norm as the sum of the squares of the two subvector norms, which are themselves induced by the respective blocks of the preconditioner. The recurrences proven in Theorem 2.1 follow directly from our ability to split the norm in this way. If one considers a more general preconditioner and considers an expression for the subvector norms, one sees quickly that the presence of cross terms prevents one from deriving recurrences such as those in Theorem 2.1.

### 3 Benefits of Residual and Error Subvector Monitoring

In the introduction, we motivated our study by mentioning that the residual subvectors in saddle-point systems (1.1) often have different physical interpretations. Therefore,
Algorithm 1: The preconditioned MINRES method with monitoring of $|\eta_{j,u}| = \|r^{(j)}\|_{p_u}$ and $|\eta_{j,p}| = \|r^{(j)}\|_{p_p}$.

**Input:** $K \in \mathbb{R}^{(m+p) \times (m+p)}$, symmetric, and $f \in \mathbb{R}^{m+p}$ as in (1.1), $P = \text{blkdiag}(P_u, P_p) \in \mathbb{R}^{(m+p) \times (m+p)}$, symmetric positive definite.

**Output:** $x^{(j)}$ for some $j$ such that $\|r^{(j)}\|_{p_p}$ satisfies some convergence criteria.

1. Set $v_0 = 0, \quad w_0 = 1$
2. Choose $x^{(0)}$, set $v_1 = \begin{bmatrix} f_u \\ f_p \end{bmatrix} - Kx^{(0)}$, $z_1 = P^{-1}v_1$
3. $\gamma_1 = \sqrt{\langle z_1, v_1 \rangle}$, $v_1 \leftarrow v_1 / \gamma_1$, $z_1 \leftarrow z_1 / \gamma_1$
4. $\psi_u = \langle z_{1,u}, v_{1,u} \rangle$, $\psi_p = \langle z_{1,p}, v_{1,p} \rangle$
5. $\mu_u = \psi_u$, $\mu_p = \psi_p$
6. $m_1 = v_1$
7. Set $\eta_0 = \gamma_1$, $\eta_{0,u} = \gamma_1 \sqrt{\psi_u}$, $\eta_{0,p} = \gamma_1 \sqrt{\psi_p}$, $s_0 = s_1 = 0$, $c_0 = c_1 = 1$
8. for $j = 1$ until convergence do
9.     $\delta_j = \langle Kz_j, z_j \rangle$
10. $v_{j+1} = Kz_j - \delta_j v_j - \gamma_j v_{j-1}$ // Lanczos
11. $z_{j+1} = P^{-1}v_{j+1}$
12. $\gamma_{j+1} = \sqrt{\langle z_{j+1}, v_{j+1} \rangle}$
13. $v_{j+1} \leftarrow v_{j+1} / \gamma_{j+1}$
14. $z_{j+1} \leftarrow z_{j+1} / \gamma_{j+1}$
15. $a_0 = c_j \delta_j - c_{j-1} s_j \gamma_j$ // Update QR factorization
16. $a_1 = \sqrt{a_0^2 + \gamma_{j+1}^2}$
17. $a_2 = s_j \delta_j + c_{j-1} c_j \gamma_j$
18. $a_3 = s_{j-1} \gamma_j$
19. $c_{j+1} = a_0 / a_1$, $s_{j+1} = \gamma_{j+1} / a_1$ // Givens rotations
20. $\theta_u \leftarrow \langle m_{j,u}, z_{j+1,u} \rangle$, $\theta_p \leftarrow \langle m_{j,p}, z_{j+1,p} \rangle$
21. $\psi_u \leftarrow \langle z_{j+1,u}, v_{j+1,u} \rangle$, $\psi_p \leftarrow \langle z_{j+1,p}, v_{j+1,p} \rangle$
22. $m_{j+1} = - s_{j+1} m_j + c_{j+1} v_{j+1}$
23. $w_{j+1} = (z_j - a_3 w_{j-1} - a_2 w_j) / a_1$
24. $x^{(j)} = x^{(j-1)} + c_j + 1 \eta_{j-1} w_{j+1}$
25. $\mu_u \leftarrow s_{j+1}^2 \mu_u - 2 s_{j+1} c_j \eta_{j-1} w_{j+1}$
26. $\mu_p \leftarrow s_{j+1}^2 \mu_p - 2 s_{j+1} c_j \eta_{j-1} w_{j+1}$
27. $\eta_j = - s_{j+1} \eta_{j-1}$ // total residual norm
28. $\eta_{j,u} = \eta_j \sqrt{\mu_u}$, $\eta_{j,p} = \eta_j \sqrt{\mu_p}$ // partial residual norms
29. Test for convergence
there is a natural interest in monitoring their norms separately. This will be illustrated in the following section by means of examples. In this section, we focus on further benefits of the ability to monitor subvector norms.

Suppose, for instance, that the user places a high value on the satisfaction of the second equation (1.1) but requires only moderate accuracy in the first equation, i.e., \( \varepsilon_p \ll \varepsilon_u \). Since the code provided at Herzog, Soodhalter, 2016 implements such refined stopping criteria

\[
|\eta_{j,u}| = \| \eta^{(j)}_u \|_{P^{-1}} \leq \varepsilon_u \quad \text{and} \quad |\eta_{j,p}| = \| \eta^{(j)}_p \|_{P^{-1}} \leq \varepsilon_p
\]

(as well as their relative counterparts), they can be directly imposed. In a traditional implementation of MINRES, the user would need to request the more stringent condition

\[
\| r^{(j)} \|_P \leq \frac{1}{\sqrt{2}} \varepsilon_p
\]

to guarantee both tolerances. This will generally lead to higher iteration numbers; see for instance Example 4.6.

Ultimately, the user may be more interested in the norm of the errors

\[
\| u - u^* \|_{P_u} \quad \text{and} \quad \| p - p^* \|_{P_p}
\]

than in the norm of the residuals. Clearly, the joint error norms could be estimated via

\[
\| u - u^* \|_{P_u}^2 + \| p - p^* \|_{P_p}^2 \leq \| K^{-1} \|_2 \| r \|_2^2
\]

where \( \| K^{-1} \|_2 = \max \{ \| K^{-1} \|_P : \| r \|_P = 1 \} \) denotes the operator norm of \( K^{-1} \). This is the approach taken in Silvester, Simoncini, 2011 to estimate the total error.

With the ability to monitor the residual subvector norms, however, it becomes feasible to estimate the two error components \( \| u - u^* \|_{P_u} \) and \( \| p - p^* \|_{P_p} \) separately. For instance, if

\[
K = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},
\]

then it follows from Krendl, Simoncini, Zulehner, 2013, Theorem 1 that the refined estimate

\[
\left( \frac{\| u - u^* \|_{P_u}}{\| p - p^* \|_{P_p}} \right) \leq \left[ \frac{1}{\alpha} \frac{\| a \|}{\beta \| a \|} \frac{\| a \|}{\beta^2 \| a \|} \right] \left( \frac{\| \eta_u \|_{P_u}}{\| \eta_p \|_{P_p}} \right)
\]

holds. Depending on the example, the positive constants

\[
\| a \| := \max_{u \neq 0} \frac{\| u^T A u \|_{P_u}}{\| u \|_{P_u}^2}, \quad \alpha := \min_{u \in \ker B \setminus \{0\}} \frac{u^T A u}{\| u \|_{P_u}^2}, \quad \beta := \min_{p \neq 0} \max_{u \neq 0} \frac{\| p^T B u \|_{P_p}}{\| u \|_{P_u} \| p \|_{P_p}}
\]

may either be available analytically or they can be estimated. Notice that due to \( K \) being invertible, \( u^T A u \) has constant sign on \( \ker B \setminus \{0\} \), and we assumed it to be positive definite for the purpose of this section.

To summarize, estimation of the residual subvector norms enables the formulation of refined stopping criteria, based on either the residual or error subvector norms. We envision numerous possible applications, for instance concerning solvers for equality constrained nonlinear optimization problems which may now assign individual stopping criteria for the residuals representing optimality and feasibility, respectively. The design of such algorithms is, however, beyond the scope of this work.
4 Examples and Numerical Experiments

In this section, we discuss several examples of saddle-point systems where the residual subvectors have different meanings. We also present numerical results which demonstrate the correctness of Algorithm 1. We do this by storing all the residual vectors throughout the iteration and evaluating the preconditioned subvector norms a-posteriori. This is a debugging step which is introduced to verify the correctness of Algorithm 1. Our MATLAB implementation of Algorithm 1 is available at Herzog, Soodhalter, 2016.

In all examples, we begin with an all-zero initial guess and we stop when the relative reduction of the total residual

\[
\frac{\eta_j}{\eta_0} = \frac{\|r^{(j)}\|_{P^{-1}}}{\|r^{(0)}\|_{P^{-1}}}
\]

falls below $10^{-6}$, unless indicated otherwise.

Example 4.1 (Under- and Overdetermined Linear Systems).

In this example we consider linear systems of the form

\[
\begin{bmatrix}
H & B^T \\
B & 0
\end{bmatrix}
\begin{bmatrix}
u \\
p
\end{bmatrix}
= \begin{bmatrix}
0 \\
b
\end{bmatrix}
\quad \text{and} \quad
\begin{bmatrix}
H & B^T \\
B & 0
\end{bmatrix}
\begin{bmatrix}
u \\
p
\end{bmatrix}
= \begin{bmatrix}
a \\
0
\end{bmatrix},
\]

(4.1)

where $H \in \mathbb{R}^{n \times n}$ is symmetric and positive definite, $B \in \mathbb{R}^{m \times n}$ is of full row rank and $m < n$ holds. The first system arises as the optimality condition of the problem

\[
\text{Minimize } \frac{1}{2} \|u\|_H^2 \quad \text{such that } \ Bu = b, \quad u \in \mathbb{R}^n
\]

of finding the least-norm solution of the underdetermined and consistent linear system $Bu = b$. Notice that the first block of the respective system in (4.1) represents optimality, while the second block represents feasibility of a candidate solution $(u, p)^T$.

Our test case uses data created through the commands

\[
\text{rng}(42); \ n=100; \ m=30; \ B=\text{randn}(m,n); \ b=\text{randn}(m,1); \\
H=\text{spdiags(rand(n,1),0,n,n)};
\]

As preconditioner we use either $P_1 = \text{blkdiag}(I_{n \times n}, I_{m \times m})$ (the unpreconditioned case) or $P_2 = \text{blkdiag}(H, I_{m \times m})$. The convergence histories in Figure 4.1 (top row) show that the amount by which the two residual subvectors contribute to their combined norm may indeed be quite different, and it depends on the preconditioner. In this example, the feasibility residual $r_p = b - Bu$ is significantly smaller than the optimality residual $r_u = -Hu - B^Tp$ in the unpreconditioned case where we used $P_1$. To be more precise, the average value of the squared residual norm fraction $\mu_p$ throughout the iteration history is about 21%; see (2.13). Quite the opposite is true for the case of $P_2$, when $\mu_p$ is close to 100%.
Next we consider an overdetermined linear system $B^T p = a$. Its least-squares solution

$$\text{Minimize } \frac{1}{2} \| B^T p - a \|_{H^{-1}}, \quad p \in \mathbb{R}^m$$

is uniquely determined by the normal equations, $B H^{-1} (B^T p - a) = 0$. By defining $u$ as the ‘preconditioned residual’ $H^{-1} (B^T p - a)$, we find that the least-squares solution is equivalent to the second saddle-point system in (4.1). The first block of equations now represents feasibility for the constraint defining the auxiliary quantity $u$, while the second requires $u$ to lie in the kernel of $B$.

Similarly as above, we derive test data through

```
rng(42); n=100; m=30; B=randn(m,n); a=randn(n,1);
H=spdiags(rand(n,1),0,n,n);
```

We employ the same preconditioners $P_1$ and $P_2$ as in the underdetermined case above. Once again, the convergence behavior of the two residual subvectors shown in Figure 4.1 (bottom row) is fundamentally different for the two cases: in the unpreconditioned setting, the average value of $\mu_p$ is about 9%, while it is 72% in the preconditioned case.

**Remark 4.2.** Figure 4.1 shows the residual subvector norms as predicted by the quantities $|\eta_j,u|$ and $|\eta_j,p|$ in Algorithm 1 as grey dots. In order to verify the correctness of our derivation and implementation, we also stored the sequence of actual residual vectors occurring during the run of Algorithm 1, and evaluated the preconditioned norms of their subvectors, i.e., $\| r^{(j)}_u \|_{P^{-1}_u}$ and $\| r^{(j)}_p \|_{P^{-1}_p}$, a-posteriori. We can observe that they coincide perfectly, which confirms the correctness of Theorem 2.1 and our implementation. It also shows that the recursive formulae for $|\eta_j,u|$ and $|\eta_j,p|$ appear to be insensitive to round-off. Traditional implementations of MINRES monitor only the total residual norm $\| r^{(j)} \|_{P^{-1}}$ (not shown). The remaining figures in this section follow this example.

Our remaining examples involve partial differential equations and they will be stated in variational form, by specifying bilinear forms $a(\cdot,\cdot) : V \times V \rightarrow \mathbb{R}$, $b(\cdot,\cdot) : V \times Q \rightarrow \mathbb{R}$, and, where appropriate, $c(\cdot,\cdot) : Q \times Q \rightarrow \mathbb{R}$. Here $V$ and $Q$ are (real) Hilbert spaces. The matrices $A$, $B$ and $C$ in (1.1) are then obtained by evaluating the respective bilinear forms on a basis of an appropriate finite dimensional subspace $V_h$ or $Q_h$, e.g.,

$$[B]_{ij} = b(\phi_j, \psi_i)$$

for basis elements $\phi_j \in V_h$ and $\psi_i \in Q_h$. Similarly, the right hand side vectors $f_u$ and $f_p$ are obtained from evaluating problem dependent linear forms $f_u(\cdot)$ and $f_p(\cdot)$ on the same basis functions $\phi_i$ and $\psi_i$, respectively. Finally, the matrices and vectors obtained in this way may need to be updated due to the incorporation of essential (Dirichlet) boundary conditions.

All numerical tests were conducted using the Python interface of the finite element library FEniCS Logg, Mardal, Wells, 2012 (version 1.6) to generate the matrices and
Figure 4.1: Convergence history of the residual subvectors for Example 4.1 (top row: underdetermined setting, bottom row: overdetermined setting) in the unpreconditioned (left) and preconditioned cases (right). See Remark 4.2 for a detailed explanation of the plots. Notice that in the top right figure, the first residual norm $\|r_u\|_{P^{-1}}$ is partially outside of the plot range. This results from our choice to maintain the same scales for both figures.

vectors. Those were then exported in PETSc binary format and read from MATLAB through the helper functions `readPetscBinMat.m` and `readPetscBinVec.m` provided by Samar Khatiwala on his web page. We deliberately turned off the reordering feature of FEniCS for the degrees of freedom to preserve the block structure of (1.1) for illustration purposes (see Figure 4.5) and in order for the subvectors $u$ and $p$ and the residuals $r_u$ and $r_p$ to remain contiguous in memory for simplicity. However, our theory does not rely on a particular ordering of the subvector components, and our implementation of Algorithm 1 allows for arbitrary component ordering.

In the subsequent examples we further illustrate our approach to calculate and monitor the residual subvector norms during the MINRES iteration. In all examples, we employ block-diagonal preconditioners which are not new but have been introduced elsewhere in the literature. All preconditioners have in common that they follow the philosophy of operator preconditioning. Briefly, this means that the preconditioners are derived
from the Hilbert space mapping properties of the differential operators giving rise to the bilinear forms \( a, b, \) and \( c \). This is the essential step towards obtaining mesh independent convergence and possibly independence of other problem parameters as well. We refer the reader to Hiptmair, 2006; Kirby, 2010; Mardal, Winther, 2011; Málek, Strakoš, 2015 for more on operator preconditioning. In addition, it will turn out to be useful to specify the physical units (Newton: \( N \), meters: \( m \), seconds: \( s \), Watt: \( W \), and Kelvin: \( K \)) for all involved quantities in the following examples.

**Example 4.3** (Stokes Channel Flow).

We consider the variational formulation of a 3D stationary Stokes channel flow configuration within the domain \( \Omega = (0, 10) \times (0, 1) \times (0, 1) \). Dirichlet conditions are imposed on the fluid velocity \( u \) everywhere except at the ‘right’ (outflow) boundary \( (x = 10) \), where do-nothing conditions hold. The Dirichlet conditions are homogeneous (no-slip) except at the ‘left’ (inflow) boundary \( (x = 0) \), where \( u(x, y, z) = u_{in}(x, y, z) = (y(1-y)z(1-z), 0, 0)^T \) \( m/s \) is imposed.

Appropriate function spaces for this setup are \( V = \{ v \in H^1(\Omega; \mathbb{R}^3) : v = 0 \text{ on } \Gamma \setminus \Gamma_{\text{right}} \} \) for the velocity and \( Q = L^2(\Omega) \) for the pressure. The relevant bilinear and linear forms associated with this problem are

\[
a(u, v) = \mu \int_\Omega \nabla u : \nabla v \, dx, \quad b(u, q) = \int_\Omega q \, \text{div} u \, dx \quad \text{and} \quad f_u(v) = \int_\Omega f \cdot v \, dx.
\]

We use the dynamic viscosity parameter \( \mu = 1 \times 10^{-3} \) \( N \cdot m^{-2} \) (water) and zero right hand side force \( f = (0, 0, 0)^T \) \( N \cdot m^{-3} \).

By considering units, or by investigating the underlying physics, we infer that the first component of the residual, \( r_u = f_u - a(u, \cdot) - b(\cdot, p) \), represents a net sum of forces, measured in N. Similarly, the second residual \( r_p = -b(u, \cdot) \) represents the net flux of fluid through the impermeable channel walls, measured in \( m^3/s \). Clearly, both parts of the residual must be zero at the converged solution, but their departure from zero at intermediate iterates has different physical interpretations.

As preconditioner \( P = \text{blkdiag}(P_u, P_p) \), we use the block diagonal matrix induced by the bilinear forms

\[
a(u, v) \quad \text{and} \quad \mu^{-1} \int_\Omega p q \, dx,
\]

respectively, similar to Elman, Silvester, Wathen, 2014, Section 4.2, where the constant-free problem is considered. Notice that the inclusion of the constant \( \mu^{-1} \) into the pressure mass matrix renders the preconditioner compatible with the physical units of the problem. Other arguments in favor of this scaling are given in ur Rehman et al., 2011.

For our numerical test, we discretized the problem using the Taylor-Hood finite element. The homogeneous and non-homogeneous Dirichlet boundary conditions were included by modifying the saddle-point components \( A, B, B^T \), and right hand side \( f_u \) in (1.1) in a symmetric way through the assemble système call in FENICS. Notice that this effectively modifies some components of the first residual subvector \( r_u \) by expressions...
of the form $u_{\text{res}}(x, y, z) - u(x, y, z)$, measured in m s$^{-1}$. The same modifications apply to the preconditioner $P$.

Figure 4.2 (top row) displays the convergence behavior of the residual subvector norms on a relatively coarse mesh and its refinement. The result illustrates the mesh independence of the preconditioned iteration, and it also shows that the residual $r_p$ (representing mass conservation) lags behind the residual $r_u$ over the majority of the iterations.

We also illustrate the estimate (3.1) for the error components. To this end, we need to determine the constants in (3.2). By construction, $\|a\| = a = 1$ holds both in the continuous and discretized settings. The inf–sup constant $\beta$ was evaluated numerically by considering the generalized, positive definite eigenvalue problem $P_p p = \lambda B P_u^{-1} B^T p$ and setting $\beta = \lambda_{\text{max}}^{-1/2}$. Notice that $\beta$ is independent of the viscosity parameter $\mu$ due to our scaling of norms. Its value was found to be approximately $\beta \approx 0.03$ on both discretization levels.

As can be seen from the bottom row of Figure 4.2, eq. (3.1) provides a rather sharp estimate of both error components over a long period of iterations but overestimates the error in the velocity and pressure by one and two orders of magnitude, for discretization levels 1 and 2, respectively, towards the end of the iteration history in this example. In any case, our estimate correctly predicts that the total error is dominated by the error in the pressure.

Example 4.4 (Linear Elasticity with Nearly Incompressible Material).

This example describes a tensile test with a rod of square cross section, which occupies the domain $\Omega = (0, 100) \times (0, 10) \times (0, 10)$. We use mm here in place of m as our length unit. Homogeneous Dirichlet conditions for the displacement $u$ are imposed at the 'left' (clamping) boundary ($x = 0$), while natural (traction) boundary conditions are imposed elsewhere. The imposed traction pressure is zero except at the 'right' (forcing) boundary ($x = 100$), where a uniform pressure of $g = (1, 0, 0)^T$ N mm$^{-2}$ is imposed. As is customary for nearly incompressible material, we introduce an extra variable $p$ for the hydrostatic pressure (see for instance Wieners, 2000) in order to overcome the ill-conditioning of a purely displacement based formulation known as locking Babuška, Suri, 1992. We employ the standard isotropic stress-strain relation, $\sigma = 2 \mu \varepsilon(u) + p I$ and $\text{div} \, u = \lambda^{-1} p$. Here $\varepsilon(u) = (\nabla u + \nabla u^T)/2$ denotes the symmetrized Jacobian of $u$, while $\mu$ and $\lambda$ denote the Lamé constants. We choose as material parameters Young’s modulus $E = 50$ N mm$^{-2}$ and a Poisson ratio of $\nu = 0.49$. These particular values describe a material like nearly incompressible rubber, and a conversion to the Lamé constants yields $\mu = \frac{E}{2(1+\nu)} = 16.78$ N mm$^{-2}$ and $\lambda = \frac{E \nu}{(1+\nu)(1-2\nu)} = 822.15$ N mm$^{-2}$.

The variational mixed formulation obtained in this way is described by the spaces $V = \{v \in H^1(\Omega; \mathbb{R}^3) : v = 0 \text{ on } \Gamma_{\text{left}}\}$ for the displacement and $Q = L^2(\Omega)$ for the pressure. 


Monitoring of Residual in Saddle-Point Problems

Herzog, Soodhalter

Figure 4.2: The top row shows the convergence history of the residual subvectors for Example 4.3 (Stokes) on a coarse grid (left) and its uniform refinement (right). The bottom row shows the evolution of the two components of the error, \( \|\mathbf{u} - \mathbf{u}^*\|_{\mathbf{P}_u} \) and \( \|\mathbf{p} - \mathbf{p}^*\|_{\mathbf{P}_p} \), as well as their estimates according to (3.1).

hydrostatic pressure. The bilinear and linear forms associated with this problem are

\[
\begin{align*}
    a(\mathbf{u}, \mathbf{v}) &= 2\mu \int_{\Omega} \varepsilon(\mathbf{u}) : \varepsilon(\mathbf{v}) \, dx, \\
    b(\mathbf{u}, q) &= \int_{\Omega} q \, \text{div} \, \mathbf{u} \, dx, \\
    c(p, q) &= \lambda^{-1} \int_{\Omega} p \, q \, dx \quad \text{and} \quad f_u(\mathbf{v}) = \int_{\Gamma_{\text{right}}} g \cdot \mathbf{v} \, dx.
\end{align*}
\]

At the converged solution, the body is in equilibrium. At an intermediate iterate, we can interpret \( r_u = f_u - a(\mathbf{u}, \cdot) - b(\cdot, \mathbf{p}) \) as a net force acting on the body, measured in N. This force is attributed to a violation of the equilibrium conditions \( \text{div} \, \sigma = -f \) in \( \Omega \), and \( \sigma \mathbf{n} = 0 \) or \( \sigma \mathbf{n} = \mathbf{g} \) at the non-clamping boundary parts at an intermediate iterate \( (\mathbf{u}, \mathbf{p})^T \). The second residual \( r_p = -b(\mathbf{u}, \cdot) + c(p, \cdot) \) admits an interpretation of a volume measured in m\(^3\) due to a violation of \( \text{div} \, \mathbf{u} = \lambda^{-1} p \).
As preconditioner $P = \text{blkdiag}(P_u, P_p)$, we use the block diagonal matrix induced by the bilinear forms $a(u, v)$ and $c(p, q)$, respectively. Once again, we remark that this choice is compatible with the physical units of the problem. We also include results for an alternative scaling of the preconditioner, using $a(u, v)$ and $\lambda c(p, q)$ as blocks. As was shown in Klawonn, 1998, this choice leads to spectral estimates for the preconditioned system which are independent of $\mu$ and $\lambda$.

Similarly as for Example 4.3, we discretize the problem using the Taylor-Hood finite element, and similar modifications due to Dirichlet displacement boundary conditions apply. Figure 4.3 illustrates the convergence behavior on successively refined meshes, revealing once again different orders of magnitude for the two components of the residual.
Example 4.5 (Optimal Boundary Control).

Our final example is an optimal boundary control problem for the stationary heat equation on the unit cube $\Omega = (0, 1) \times (0, 1) \times (0, 1)$ with boundary $\Gamma$. The problem statement is

\[
\text{Minimize } \frac{\alpha_1}{2} \| u - u_d \|_{L^2(\Omega)}^2 + \frac{\alpha_2}{2} \| f \|_{L^2(\Gamma)}^2 \\
\text{s.t. } \begin{cases}
-\kappa \Delta u + \delta u &= 0 \quad \text{in } \Omega, \\
\kappa \frac{\partial u}{\partial n} &= f \quad \text{on } \Gamma.
\end{cases}
\]

As data we use $\alpha_1 = 1 \text{ m}^{-3} \text{K}^{-2}$, $u_d(x,y,z) = x \text{ K}$, $\alpha_2 = 1 \times 10^{-2} \text{m}^2 \text{W}^{-2}$, heat conduction coefficient $\kappa = 1 \text{ W m}^{-1} \text{K}^{-1}$, and radiation coefficient $\delta = 1 \text{ W m}^{-3} \text{K}^{-1}$.

The necessary and sufficient optimality conditions for this problem are standard; see, for instance, Tröltzsch, 2010, Chapter 2.8. When we eliminate the control function $f$ from the problem, a saddle-point system for the state $u$ and adjoint state $p$ remains, which is described by the following data:

\[
a(u,v) = \alpha_1 \int_{\Omega} u v \, dx, \quad b(u,q) = \kappa \int_{\Omega} \nabla u \cdot \nabla q \, dx + \delta \int_{\Omega} u q \, dx, \\
c(p,q) = \alpha_2^{-1} \int_{\Omega} p q \, dx \quad \text{and} \quad f_u(v) = \alpha_1 \int_{\Omega} u_d v \, dx.
\]

The underlying spaces are $V = Q = H^1(\Omega)$, and we discretize both using piecewise linear, continuous finite elements. Note that the elements $u,v \in V$ are measured in K while the elements $p,q \in Q$ are measured in W$^{-1}$.

As preconditioner $P = \text{blkdiag}(P_u,P_p)$, we use the block diagonal matrix induced by the bilinear forms

\[
p_u(u,v) = \alpha_1 \kappa \delta^{-1} \int_{\Omega} \nabla u \cdot \nabla v \, dx + \alpha_1 \int_{\Omega} u v \, dx \\
p_p(p,q) = \alpha_2^{-1}(\kappa \delta^{-1})^{1/2} \int_{\Omega} \nabla p \cdot \nabla q \, dx + \alpha_2^{-1}(\kappa^{-1} \delta)^{1/2} \int_{\Omega} p q \, dx,
\]

respectively. As in our previous examples, this choice is compatible with the physical units of the problem. Figure 4.4 illustrates the convergence behavior on successively refined meshes. Besides the mesh independence, we observe that both residual subvector norms converge in unison in this example.

Example 4.6 (Image Deblurring).

Our final example is a simple constrained, regularized image deblurring problem; compare, e.g., Bredies, Lorenz, 2011, Example 6.97. We consider the following minimization problem,

\[
\text{Minimize } \frac{1}{2} \| H u - f \|_2^2 + \frac{\mu}{2} \| D_1 u \|_2^2 + \frac{\mu}{2} \| D_2 u \|_2^2, \quad u \in \mathbb{R}^{n_1 \cdot n_2} \\
\text{s.t. } B u = B f.
\] (4.2)
Here \( n_1 \) and \( n_2 \) are the dimensions (number of pixels) of the image \( u \) to be recovered, and \( H \) is a discrete Gaussian blur operator implemented by the following code utilizing the MATLAB Imaging Toolbox,

\[
\begin{align*}
  h &= \text{fspecial}(\text{`gaussian'},[5 5],2); \\
  H &= @(im) \text{imfilter}(im,h,\text{`symmetric','conv'});
\end{align*}
\]

Moreover, the data \( f \) is obtained from the original image (see Figure 4.7) by applying the blur matrix \( H \) and white noise with standard deviation \( \sigma = 30 \):

\[
\begin{align*}
  \text{im\_blurred} &= H(\text{double(im)}); \\
  \text{im\_blurred\_noisy} &= \text{im\_blurred} + 30*\text{randn(size(im\_blurred))};
\end{align*}
\]

We apply a simple isotropic diffusion-based regularization with parameter \( \mu > 0 \), i.e., \( D_1 \in \mathbb{R}^{n_2 \times (n_1-1) \times n_1 \times n_2} \) and \( D_2 \in \mathbb{R}^{n_1 \times (n_2-1) \times n_1 \times n_2} \) are horizontal and vertical difference operators. We found \( \mu = 1.8 \) to be a suitable value for the given noise level. Finally the constraint \( Bu = Bf \) represents preservation of total brightness, i.e., \( B = (1, \ldots, 1) \in \mathbb{R}^{1 \times n_1 \times n_2} \) holds. Since \( H \) is a symmetric matrix, the problem gives rise to the following saddle-point problem as its optimality conditions,

\[
\begin{bmatrix}
  A & B^T \\
  B & 0
\end{bmatrix}
\begin{bmatrix}
  u \\
  p
\end{bmatrix} =
\begin{bmatrix}
  f \\
  Bf
\end{bmatrix},
\]

where \( A = H^2 + \mu D_1^T D_1 + \mu D_2^T D_2 \).

In Figure 4.7 we present results for unpreconditioned iterations of our method, applied to the cameraman test image of size \( n_1 = n_2 = 256 \) and with image data in the range \([0, 255]\). Due to the presence of noise in the observed image data \( f \), it is not important
to obtain a very small first residual $r_u = f - Au - B^T p$ to obtain good reconstructions. We therefore applied the absolute stopping criterion

$$\|r_u^{(j)}\|_2 \leq \varepsilon_u$$

on the first residual subvector starting with a relatively coarse tolerance of $\varepsilon_u = 10^3$. We then repeated the experiment and reduced this tolerance gradually down to $10^{-6}$. As for the second residual subvector $r_p = Bf - Bu$ pertaining to the preservation of the total brightness constraint, we used

$$\|r_p^{(j)}\|_2 = |r_p^{(j)}| \leq \varepsilon_p = 10^{-2}$$

throughout, which is relatively tight in view of the value of $Bf$ being in the order of $7.8 \cdot 10^6$. Both stopping criteria needed to be satisfied.

The approximation of the minimizer of Tikhonov functionals in the unconstrained case has been treated in the ill-posed problems context previously; see, e.g., Neubauer, 1989, in which for the more general (and complicated) nonlinear case, error estimates and tolerance prescriptions are given to ensure that such a method remains (in the analysis sense) a regularization. In this particular case, one can do the same based on the fact that for $\mu > 0$, the Tikhonov functional is convex and the induced operator well-posed. Further discussion is beyond the scope of this paper, and we refer the reader to Engl, Hanke, Neubauer, 1996 for more general discussions on this topic.

In this example we clearly benefit from the ability to specify individual tolerances for both residual subvectors. Since the data $f$ is subject to significant noise, we can content ourselves with a relatively coarse residual tolerance on the first subvector. It turns out that the results remain qualitatively the same for $\varepsilon_u = 10^2$ and below; see Figure 4.6. On the other hand, if brightness preservation of the observed data is a concern, then a relatively tight tolerance of $\varepsilon_p = 10^{-2}$ can be set for the second residual subvector and it does not impair fast convergence; MINRES requires only 13 iterations to converge to $(\varepsilon_u, \varepsilon_p) = (10^2, 10^{-2})$. Without subvector residual monitoring, one would have to impose the tighter stopping criterion $\|r^{(j)}\|_2 \leq \left(\frac{1}{\sqrt{2}}\right) \cdot 10^{-2}$ for the total residual vector (see the discussion in the beginning of Section 3), which would lead to 33 iterations.

5 DISCUSSION

In this paper we developed a modified implementation of MINRES. When applied to saddle-point systems, the new implementation allows us to monitor the norms of the subvectors $\|r_u^{(j)}\|_{p_u^{-1}}$ and $\|r_p^{(j)}\|_{p_p^{-1}}$ individually, while conventional implementations keep track of only the total residual norm $\|r^{(j)}\|_{p^{-1}}$. It should be obvious how Algorithm 1 generalizes to systems with more than two residual subvectors and block-
diagonal preconditioners structured accordingly. The price to pay to monitor the sub-
vector norms is the storage of one additional vector for components $m_{j,u}$ and $m_{j,p}$ com-
pared to the implementation of MINRES given in Elman, Silvester, Wathen, 2014, Al-
gorithm 4.1 as well as some additional inner product calculations. It should be noted 
that if the cost of applying the operator and preconditioner are cheap enough, it may be 
preferable to simply construct the full residual and calculate preconditioned subvector 
 norms explicitly. However, in the case that application of these operators has signifi-
cant costs (particularly in the case that they are available only as procedures and never 
actually constructed) the method presented in this paper has important advantages.

While we developed the details in finite dimensions using matrices and vectors, our 
 approach directly transfers to linear saddle-point systems in a Hilbert space setting 
using linear operators and linear forms, as described in G"unnell, Herzog, Sachs, 2014.

Being able to differentiate between the contributions to the total residual offers new 
opportunities for the design of iterative algorithms for nonlinear problems, which 
require inexact solves of (1.1) as their main ingredient. We envision for instance solvers 
for equality constrained nonlinear optimization problems which may now assign in-
dividual stopping criteria for the residuals representing optimality and feasibility, re-
spectively. The design of such an algorithm is, however, beyond the scope of this work. 
Similarly, for the Stokes Example 4.3, the user may now assign individual stopping cri-
tera for the fulfillment of the balance of forces (first residual) and the conservation of 
mass (second residual).

We envision another class of applications in constrained least-squares problems, where 
measurement noise affects only one part of the residual vector and therefore reasonable 
tolerances for this part may differ from tolerances for other parts. Exploiting this can 
lead to significantly lower iteration numbers, as demonstrated in Example 4.6. Finally, 
the ability to monitor the residual subvector norms allows a refined estimation also for 
the subvector norms of the error in saddle-point problems, as detailed in (3.1). Never-
theless we mention that error estimates based on (partial) operator norms such as (3.1) 
may be pessimistic in practice.

A topic of future research could be to study the decay properties of the subvector norms, 
rather than study the decay of the total residual norm; see for instance Elman, Silvester, 
Wathen, 2014, Section 4.2.4. We expect that such an analysis will be more involved but 
it may shed light on how the relative scaling of the preconditioner blocks affects the 
convergence of the residual subvectors.

Acknowledgment

The authors would like to thank David Silvester and Valeria Simoncini as well as Do-
iminique Orban for their comments on an earlier version of the manuscript. They also 
thank three anonymous reviewers for their suggestions which helped improve the pa-
er.


References


Figure 4.5: Sparsity plots of the saddle-point systems arising in Example 4.1 (under- and overdetermined linear systems), Example 4.3 (3D Stokes), Example 4.4 (3D elasticity), and Example 4.5 (3D optimal boundary control).
Figure 4.6: Original image (top left) with blur (top middle) and additional noise (top right). Bottom row: solution of the deblurring problem (4.2) in Example 4.6 for tolerances $\varepsilon_u \in \{10^3, 10^2, 10^{-6}\}$ and $\varepsilon_p = 10^{-2}$. MINRES requires 5/13/53 iterations to converge to these tolerances; see Figure 4.7.

Figure 4.7: Convergence history of the residual subvectors for Example 4.6 (image deblurring).