R3GMRES: including prior information in GMRES-type methods for discrete inverse problems

Dong, Yiqiu; Garde, Henrik; Hansen, Per Christian

Published in:
Electronic Transactions on Numerical Analysis

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Electronic Transactions on Numerical Analysis.
Volume 42, pp. 136-146, 2014.

ISSN 1068-9613.

ETNA
Kent State University
http://etna.math.kent.edu

R^{3}GMRES: INCLUDING PRIOR INFORMATION IN GMRES-TYPE METHODS FOR DISCRETE INVERSE PROBLEMS*

YIQIU DONG†, HENRIK GARDE†, AND PER CHRISTIAN HANSEN†

Dedicated to Lothar Reichel on the occasion of his 60th birthday

Abstract. Lothar Reichel and his collaborators proposed several iterative algorithms that augment the underlying Krylov subspace with an additional low-dimensional subspace in order to produce improved regularized solutions. We take a closer look at this approach and investigate a particular Regularized Range-Restricted GMRES method, R^{3}GMRES, with a subspace that represents prior information about the solution. We discuss the implementation of this approach and demonstrate its advantage by means of several test problems.

Key words. inverse problems, regularizing iterations, large-scale problems, prior information

AMS subject classifications. 65F22, 65F10

1. Introduction. This paper deals with iterative Krylov subspace methods for solving large ill-conditioned systems of linear equations arising from the discretization of inverse problems. Lothar Reichel has made numerous contributions in this area (as we shall see below) on which the present work builds. We consider discrete inverse problems of the form

$$\min_x \|Ax - b\|_2^2, \qquad A \in \mathbb{R}^{n\times n}, \quad b, x \in \mathbb{R}^n, \tag{1.1}$$

and we note that this problem takes the form $Ax = b$ when the square matrix A has full rank. To compute a stable solution to this problem, one must incorporate prior information about the desired solution. Often this information takes the form of a requirement concerning the smoothness of the solution, but the information can also be specified in the form of a low-dimensional "signal subspace" in which the solution must lie; cf. [11].

The latter approach is particularly attractive for large-scale problems, where the signal subspace can take the form of a Krylov subspace such as

$\mathcal{K}_j(A^\top A, A^\top b)$ for the CGLS and LSQR algorithms [11, 20],
$\mathcal{K}_j(A, b)$ for the GMRES and MINRES algorithms [5, 16],
$\mathcal{K}_j(A, Ab)$ for the RRGMRES and MR-II algorithms [4, 8, 18],

where $\mathcal{K}_j(M, v) \equiv \text{span}\{v, Mv, M^2v, \ldots, M^{j-1}v\}$ and j is the number of iterations. Depending on the application, one or more of these subspaces may be well suited to compute a good regularized solution, i.e., a good approximation that is only little sensitive to perturbations of the data; cf. [15]. Moreover, it is possible to "subspace precondition" these methods in order to favorably adjust the above Krylov subspaces if needed; cf. [7, 9, 13].

We can further improve the regularized solution by incorporating additional specific prior information. For example, we may know that the solution has a significant component in a given subspace $\mathcal{W}_p$ of dimension $p \ll j$ (e.g., chosen to represent known smoothness properties or known discontinuities). In connection with the above Krylov subspace methods, Reichel and his collaborators [1, 2, 3, 6] therefore proposed to decompose the solution into a component in $\mathcal{W}_p$ and another component in the orthogonal complement $\mathcal{W}_p^\perp$, which leads to the idea of augmented Krylov subspace methods; see also [17].

*Received September 22, 2013. Accepted July 28, 2014. Published online on September 5, 2014. Recommended by Hassane Sadok. The authors are supported by Advanced Grant No. 291405 from the European Research Council.
†Department of Applied Mathematics and Computer Science, Technical University of Denmark ({yido, hgar, pcha}@dtu.dk).

ETNA
Kent State University
http://etna.math.kent.edu

This work focuses on the range-restricted GMRES (RRGMRES) method, which was designed for rank-deficient inconsistent systems [4] and which performs better than GMRES under the influence of noisy data [15]. We consider a particular augmentation approach in which we compute regularized solutions in a signal subspace $\mathcal{S}_{p,j}$ that is the direct sum of the two subspaces $\mathcal{W}_p$ and $\mathcal{K}_j(A, Ab)$,

$$\mathcal{S}_{p,j} = \mathcal{W}_p + \mathcal{K}_j(A, Ab) \equiv \{y + z \,|\, y \in \mathcal{W}_p \wedge z \in \mathcal{K}_j(A, Ab)\}, \tag{1.2}$$

which itself is a linear subspace. In particular, we discuss how to implement the associated algorithm Regularized RRGMRES (R^{3}GMRES) efficiently, and we demonstrate its usefulness on selected test problems—including comparisons with related algorithms. Analogous implementation issues related to CGLS and MR-II are left to future work.

In Section 2 we summarize the decomposition approach and the associated augmented RRGMRES method, and we argue why a different approach is needed to compute regularized solutions in the subspace $\mathcal{S}_{p,j}$. In Section 3 we discuss the implementation details of our algorithm, and we present several numerical examples in Section 4. Conclusions are drawn in Section 5.

2. Incorporating prior information in regularizing iterations. The idea of incorporating prior information about the solution is at the heart of all regularization methods. For example, in the Tikhonov problem

$$\min_x \left\{ \|Ax - b\|_2^2 + \lambda^2 \|Lx\|_2^2 \right\}, \tag{2.1}$$

we explicitly require that the solution has a small (semi-)norm as measured by the term $\|Lx\|_2$. The matrix L is often chosen as a discrete approximation to a differential operator (to enforce smoothness of the solution), and we can modify L to incorporate other known features into the solution.

As an example, if we wish to allow a discontinuity between the solution elements x_ℓ and $x_{\ell+1}$ for $1 \le \ell \le n-1$, we can define the subspace $\mathcal{W}_2$ by

$$\mathcal{W}_2 = \text{span}\{w_1, w_2\}, \quad w_1 = \begin{bmatrix} \text{ones}(\ell, 1) \\ \text{zeros}(n-\ell, 1) \end{bmatrix}, \quad w_2 = \begin{bmatrix} \text{zeros}(\ell, 1) \\ \text{ones}(n-\ell, 1) \end{bmatrix}. \tag{2.2}$$

If the columns of W_2 form an orthonormal basis for $\mathcal{W}_2$, then $P_{W_2}^\perp = I - W_2 W_2^\top$ is the orthonormal projector on $\mathcal{W}_2^\perp$. Hence, any linear combination of w_1 and w_2 is in the null space of $LP_{W_2}^\perp$. Substituting $\|LP_{W_2}^\perp x\|_2$ for $\|Lx\|_2$ in (2.1) therefore ensures that any piecewise constant solution with the desired breakpoint is not affected by the regularization.

This idea immediately generalizes to a general subspace $\mathcal{W}_p$ and the associated projectors $P_{W_p} = W_p W_p^\top$ and $P_{W_p}^\perp = I - W_p W_p^\top$, where range$(W_p) = \mathcal{W}_p$ and W_p has orthonormal columns. Moreover, the idea carries over to the subspace preconditioned versions of the CGLS, LSQR, RRGMRES, and MR-II algorithms, and implementations such as those in Regularization Tools [10] can be used whenever it is feasible to perform operations with the oblique pseudoinverse of $LP_{W_p}^\perp$ [12]. When it is impractical to perform these operations, the approach by Hochstenbach and Reichel [14] can be used.

2.1. The decomposition approach and the augmented Krylov subspace method. The principle of leaving the solution component in $\mathcal{W}_p$ unaffected by the regularization is key to many regularization methods, and it also underlies the decomposition method in [1], which splits the solution space into a Krylov subspace that is determined by the iterative method (such as GMRES, RRGMRES, or LSQR) and the auxiliary subspace $\mathcal{W}_p$ mentioned

ETNA
Kent State University
http://etna.math.kent.edu

before. Let the span of the orthonormal columns of W_p represent the subspace $\mathcal{W}_p$. Then the computed approximate solution $x^{(j)}$, for $j = 1, 2, 3, \ldots$, is partitioned as

$$x^{(j)} = \hat{x}^{(j)} + \tilde{x}^{(j)}, \qquad \hat{x}^{(j)} = P_{W_p} x^{(j)}, \qquad \tilde{x}^{(j)} = P_{W_p}^{\perp} x^{(j)}.$$

Since the dimension p of $\mathcal{W}_p$ is assumed to be small, the component $\hat{x}^{(j)}$ is determined by solving a small linear system of equations by a direct method, while the component $\tilde{x}^{(j)}$ is computed by the GMRES or RRGMRES iterative method.

This decomposition method based on GMRES (RRGMRES) is, in fact, *equivalent* to the augmented GMRES (RRGMRES) method described in [2]; see [1, Theorem 2.2]. In the augmented method, a Krylov subspace generated by GMRES (RRGMRES) is augmented by the space $\mathcal{W}_p$ in order to make full use of the prior information. Following [2], we introduce the QR factorization

$$AW_p = V_p R, \tag{2.3}$$

where $V_p \in \mathbb{R}^{n \times p}$ has orthonormal columns and $R \in \mathbb{R}^{p \times p}$ is upper triangular. Instead of using the standard implementation of GMRES (RRGMRES) based on the Arnoldi process, in [2] the approximate solution $x^{(j)}$ of (1.1) is determined by solving the constrained least squares problem

$$\min_x \|Ax - b\|_2^2 \qquad \text{s.t.} \qquad x \in \mathcal{W}_p + \mathcal{K}_j(P_{V_p}^{\perp} A, u),$$

where $u = P_{V_p}^{\perp} b$ for augmented GMRES and $u = P_{V_p}^{\perp} Ab$ for augmented RRGMRES. Specifically, after j steps of the modified Arnoldi process, $x^{(j)}$ is computed via the modified Arnoldi decomposition

$$A \begin{bmatrix} W_p & \bar{V}_{p+1:p+j} \end{bmatrix} = \bar{V}_{p+j+1} \bar{H}_{p+j}. \tag{2.4}$$

Here, $\bar{H}_{p+j} \in \mathbb{R}^{(p+j+1) \times (p+j)}$ is an upper Hessenberg matrix whose leading principal $p \times p$ submatrix is R from (2.3). The matrix $\bar{V}_{p+j+1} = \begin{bmatrix} V_p & \bar{V}_{p+1:p+j} & \bar{v}_{p+j+1} \end{bmatrix} \in \mathbb{R}^{n \times (p+j+1)}$ has orthonormal columns, and the first column $\bar{v}_{p+1}$ of $\bar{V}_{p+1:p+j}$ is given by

$$\bar{v}_{p+1} = \begin{cases} P_{V_p}^{\perp} b / \|P_{V_p}^{\perp} b\|_2 & \text{for augmented GMRES,} \\ P_{V_p}^{\perp} Ab / \|P_{V_p}^{\perp} Ab\|_2 & \text{for augmented RRGMRES.} \end{cases}$$

Then, the iterate $x^{(j)}$ can be expressed as

$$x^{(j)} = \begin{bmatrix} W_p & \bar{V}_{p+1:p+j} \end{bmatrix} y^{(j)},$$

where $y^{(j)} \in \mathbb{R}^{p+j}$ solves the least squares problem

$$\min_y \|\bar{H}_{p+j} y - \bar{V}_{p+j+1}^{\top} b\|_2^2. \tag{2.5}$$

REMARK 2.1. The augmented method in [2] and the equivalent decomposition method in [1] both use a modified Arnoldi process that produces orthonormal vectors which are orthogonal to the columns of V_p. The basis generated by this approach corresponds to a Krylov subspace limited to the orthogonal complement of $\mathcal{V}_p = \text{range}(V_p)$.

In other words, the generated approximate solution $x^{(j)}$ in the jth iteration lies in an augmentation of the Krylov subspace $\mathcal{K}_j(P_{V_p}^{\perp} A, P_{V_p}^{\perp} b)$ for GMRES and $\mathcal{K}_j(P_{V_p}^{\perp} A, P_{V_p}^{\perp} Ab)$ for RRGMRES, instead of $\mathcal{K}_j(A, b)$ and $\mathcal{K}_j(A, Ab)$, respectively, as one would expect.

ETNA
Kent State University
http://etna.math.kent.edu

2.2. Another augmented Krylov subspace method. In this work we want to solve the least-squares problem (1.1) in the subspace $\mathcal{S}_{p,j}$ (1.2), so we should restrict the Krylov subspace to the orthogonal complement of $\mathcal{W}_p$ instead of $\mathcal{V}_p$. To understand the relation between $\mathcal{W}_p$ and $\mathcal{V}_p$, we have the following result.

PROPOSITION 2.2. *Assume that $A \in \mathbb{R}^{n\times n}$ is nonsingular with the QR factorization of AW_p given in (2.3). The subspaces $\mathcal{W}_p$ and $\mathcal{V}_p$ are spanned by the columns of the matrices W_p and V_p, respectively. Then, $\mathcal{V}_p = \mathcal{A}(\mathcal{W}_p)$, where $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^n$ is the linear operator defined as $\mathcal{A}(w) = Aw$ with $w \in \mathcal{W}_p$.*

For the rest of the paper we focus on the RRGMRES method. The Cayley-Hamilton theorem [19] states that the inverse of a matrix can be formed as a linear combination of its powers. In order to obtain higher accuracy, we therefore prefer an approximate solution of (1.1) in its Krylov subspace, i.e., $\mathcal{K}_j(A, Ab)$ instead of $\mathcal{K}_j(P_{V_p}^{\perp}A, P_{V_p}^{\perp}Ab)$. For example, if the exact solution x^* to (1.1) is in the subspace $\mathcal{W}_p^{\perp} \cap \mathcal{V}_p$, then the approximate solution obtained by solving the least-squares problem (1.1) in $\mathcal{S}_{p,j}$ could provide higher accuracy than the one in $\mathcal{W}_p + \mathcal{K}_j(P_{V_p}^{\perp}A, P_{V_p}^{\perp}Ab)$. Below we formulate a simple extension of the augmented RRGMRES method proposed in [2] to solve (1.1) in $\mathcal{S}_{p,j}$.

In order to ensure that the approximate solution is in $\mathcal{S}_{p,j}$, the intuitive way to extend the augmented RRGMRES method is to find a decomposition of the form

$$A\begin{bmatrix} W_p & Ab & A^2b & \dots & A^jb \end{bmatrix} = V_{p+j+1}H_{p+j}, \tag{2.6}$$

which is similar to the modified Arnoldi decomposition in (2.4). Then the iterate $x^{(j)}$ can be expressed as

$$x^{(j)} = \begin{bmatrix} W_p & Ab & A^2b & \dots & A^jb \end{bmatrix} y^{(j)},$$

where $\{Ab, A^2b, \cdots, A^jb\}$ forms a basis of $\mathcal{K}_j(A, Ab)$, and $y^{(j)}$ solves the same least squares problem as in (2.5).

From a numerical point of view, the "naive" basis $\{Ab, A^2b, \dots, A^jb\}$ of the Krylov subspace $\mathcal{K}_j(A, Ab)$ is not a good choice. As j increases, most of the vectors in this basis will point more and more into the same direction. Thus, this basis is usually ill-conditioned, which leads to a severe loss of precision and even breakdown after some iterations. Hence, in the algorithm below we apply a Modified Gram-Schmidt (MGS) orthonormalization to the basis $\{Ab, A^2b, \dots, A^jb\}$. See also the discussion of implementation issues in [18].

ALGORITHM 1. INTUITIVE VERSION.

1: Compute the QR factorization $AW_p = V_pH_p$, where $V_p \in \mathbb{R}^{n\times p}$ and $H_p \in \mathbb{R}^{p\times p}$.
2: Let $u_1 = Ab$, $v_{p+1} = P_{V_p}^{\perp}u_1$, and normalize $u_1 = u_1/\|u_1\|_2$, $v_{p+1} = v_{p+1}/\|v_{p+1}\|_2$. Then, expand $V_{p+1} := \begin{bmatrix} V_p & v_{p+1} \end{bmatrix}$ and $W_{p+1} := \begin{bmatrix} W_p & u_1 \end{bmatrix}$.
3: Initialize $R_1 := 1$, and set $j := 1$.
4: Compute $v_{p+j+1} = Au_j$ and $u_{j+1} = v_{p+j+1}/\|v_{p+j+1}\|_2$.
5: Apply MGS orthonormalization to v_{p+j+1}, and expand $V_{p+j+1} := [V_{p+j} \;\; v_{p+j+1}]$, $H_{p+j} := \begin{bmatrix} H_{p+j-1} & h_{p+j} \\ 0 & \end{bmatrix} \in \mathbb{R}^{(p+j+1)\times(p+j)}$, where h_{p+j} is from the MGS.
6: Solve

$$\min_y \left\| H_{p+j} \begin{bmatrix} I_p & 0 \\ 0 & R_j^{-1} \end{bmatrix} y - V_{p+j+1}^{\top}b \right\|_2^2$$

to obtain $y^{(j)}$. Then $x^{(j)} = W_{p+j}y^{(j)}$.

ETNA
Kent State University
http://etna.math.kent.edu

7: Apply MGS orthonormalization to u_{j+1} such that $\{u_1, \ldots, u_{j+1}\}$ becomes an orthonormal basis for $\mathcal{K}_{j+1}(A, Ab)$, expand $W_{p+j+1} = [\, W_{p+1} \;\; u_{j+1} \,]$, and expand $R_{j+1} := \begin{bmatrix} R_j & r_{j+1} \\ 0 & \end{bmatrix} \in \mathbb{R}^{(j+1)\times(j+1)}$, where r_{j+1} is from the MGS.
8: Stop, or set $j := j + 1$, and return to step 4.

Because of the extra MGS orthonormalization in step 7, the work in Algorithm 1 is dominated by two MGS processes, and the algorithm therefore asymptotically needs additional $O(j^2 n)$ flops compared to the augmented RRGMRES method proposed in [2]. In order to find a more efficient algorithm, in the next section we take into account the low dimension of $\mathcal{W}_p$ and reorganize the framework of the augmented RRGMRES method.

3. Implementation of the Regularized RRGMRES (R^{3}GMRES) method. In the algorithm described below, instead of appending $\mathcal{W}_p$ by the basis of $\mathcal{K}_j(A, Ab)$ as in (2.6), we use the standard Arnoldi process to determine an orthonormal basis of $\mathcal{K}_j(A, Ab)$ and then augment it by $\mathcal{W}_p$ in each step of the iterative algorithm. While this may seem cumbersome, we shall see that the computational overhead is favorably smaller than that of Algorithm 1.

3.1. The basic algorithm. Our new Regularized RRGMRES method, R^{3}GMRES, is based on the decomposition

$$A\,[\, V_j \;\; W_p \,] = \begin{bmatrix} V_{j+1} & \widetilde{V}_j \end{bmatrix} \begin{bmatrix} H_j & G_j \\ 0 & F_j \end{bmatrix}, \tag{3.1}$$

where $AV_j = V_{j+1}H_j$ is obtained after j steps of the Arnoldi process. Here, $V_j \in \mathbb{R}^{n\times j}$ has orthonormal columns with the first column $v_1 = Ab/\|Ab\|_2$, and $H_j \in \mathbb{R}^{(j+1)\times j}$ is an upper Hessenberg matrix. The columns of V_j form an orthonormal basis of the Krylov subspace $\mathcal{K}_j(A, Ab)$. We then augment this basis to a basis of $\mathcal{S}_{p,j}$, which turns out to be the augmented matrix $[\, V_j \;\; W_p \,]$.

We must also augment V_{j+1} with a basis of the range of AW_p, which leads to the augmented matrix $[\, V_{j+1} \;\; \widetilde{V}_j \,]$, where the orthonormal vectors in $\widetilde{V}_j \in \mathbb{R}^{n\times p}$ are also orthogonal to the columns of V_{j+1}. Furthermore, we introduce the two matrices $G_j \in \mathbb{R}^{(j+1)\times p}$ and $F_j \in \mathbb{R}^{p\times p}$ which are composed of the coefficients of AW_p with respect to the basis of $\mathcal{V}_{j+1}$ and the subspace of $\mathcal{V}_{j+1}^{\perp}$, respectively:

$$G_j = V_{j+1}^{\top} A W_p, \qquad F_j = \widetilde{V}_j^{\top} A W_p.$$

Substituting (3.1) into (1.1) shows that the iterate $x^{(j)} \in \mathcal{S}_{p,j}$ can be expressed as

$$x^{(j)} = [\, V_j \;\; W_p \,]\, y^{(j)},$$

where $y^{(j)}$ solves the least squares problem

$$\min_y \left\| \begin{bmatrix} H_j & G_j \\ 0 & F_j \end{bmatrix} y - \begin{bmatrix} V_{j+1}^{\top} \\ \widetilde{V}_j^{\top} \end{bmatrix} b \right\|_2^2. \tag{3.2}$$

This leads to the R^{3}GMRES method presented below.

ETNA
Kent State University
http://etna.math.kent.edu

ALGORITHM 2. R^{3}GMRES METHOD.

1: Set $v_1 = Ab/\|Ab\|_2$, $V_1 := v_1$, $G_0 := v_1^\top AW_p$, and $j := 1$.
2: Use the Arnoldi process to obtain v_{j+1} and h_j such that $AV_j = V_{j+1}H_j$, where $V_{j+1} := [\,V_j \;\; v_{j+1}\,]$ and $H_j := \begin{bmatrix} H_{j-1} & h_j \\ 0 & \end{bmatrix} \in \mathbb{R}^{(j+1)\times j}$ (with $H_1 = h_1$).
3: Compute $G_j = \begin{bmatrix} G_{j-1} \\ v_{j+1}^\top AW_p \end{bmatrix} \in \mathbb{R}^{(j+1)\times p}$.
4: Orthonormalize AW_p with respect to V_{j+1} to obtain $\widetilde{V}_j$, and compute $F_j = \widetilde{V}_j^\top AW_p$.
5: Solve (3.2) to obtain $y^{(j)}$. Then $x^{(j)} = [\,V_j \;\; W_p\,]\, y^{(j)}$.
6: Stop, or set $j := j + 1$, and return to step 2.

In every iteration we need to recompute $\widetilde{V}_j$ and F_j in step 4 because of the expansion of the Krylov subspace $\mathcal{K}_j(A, Ab)$, but due to the small value of p, the computational work of Algorithm 2 is still dominated by the Arnoldi process (see the implementation details below). Hence, the computational work of the new R^{3}GMRES method is asymptotically the same as that of the augmented RRGMRES method in [2].

3.2. Implementation details. The key to an efficient implementation is to update an orthogonal factorization of the coefficient matrix in the least squares problem (3.2)

$$\begin{bmatrix} H_j & G_j \\ 0 & F_j \end{bmatrix} = Q \begin{bmatrix} T_j^{(11)} & T_j^{(12)} \\ 0 & T_j^{(22)} \\ 0 & 0 \end{bmatrix},$$

where $T_j^{(11)} \in \mathbb{R}^{j\times j}$ and $T_j^{(22)} \in \mathbb{R}^{p\times p}$ are upper triangular and Q is orthogonal.

The submatrix $T_j^{(11)}$ is updated via Givens transformations as in the standard GMRES and RRGMRES algorithms; the rotations are also applied to G_j and the right-hand side, i.e., to $V_{j+1}^\top b$. At this stage, before treating the submatrix F_j, we have an intermediate system of the form (shown for $j = 3$ and $p = 2$):

$$\begin{bmatrix} T_j^{(11)} & \text{intermediate} \\ 0 & F_j \quad \widetilde{V}_j^\top b \end{bmatrix} = \left[\begin{array}{ccc|cc||c} \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ \hline & & & \times & \times & \times \\ & & & \times & \times & \times \end{array}\right] \leftarrow \text{store row } j+1,$$

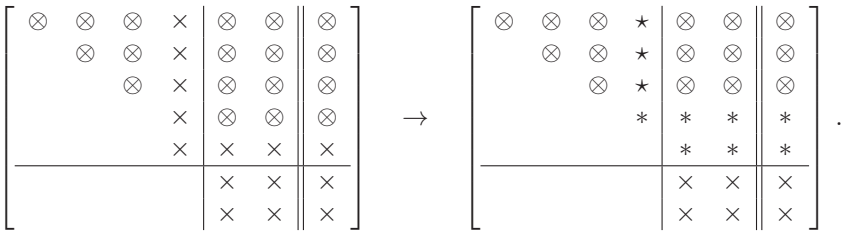
where the rightmost column represents the right-hand side. We need to store row $j + 1$ of the intermediate system for reasons that will be explained below. To complete the orthogonal reduction, we apply an orthogonal transformation that involves the bottom $p + 1$ rows of the system and produces a system of the form

$$\left[\begin{array}{ccc|cc||c} \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & * & * & * \\ \hline & & & & * & * \\ & & & & & * \end{array}\right],$$

where $*$ denotes an element that has changed. Note that $T_j^{(22)}$ in this example consists of the elements in rows 4–5 and columns 4–5.

ETNA
Kent State University
http://etna.math.kent.edu

In the next iteration (here $j = 4$), the Arnoldi process produces a new column of H_j that is inserted as column j in the (1,1)-block. This block is then reduced to upper triangular form



The elements denoted by $\otimes$ are from the intermediate system of the previous iteration, while those denote by $\times$ are new. After the reduction, the elements denoted by $\star$ are updated by means of the stored Givens transformation from the previous iterations, and those denoted by $*$ are transformed by the new Givens rotation. This is followed by an orthogonal transformation involving the bottom $p + 1$ rows of the system as before.

The key observation here is that in the previous iteration ($j = 3$), row $j + 1 = 4$ of the intermediate system was *overwritten* (to obtain triangular form). Therefore, we must store this row, so that we can insert it again into the system at the beginning of the next iteration ($j = 4$) before the Givens rotation is applied.

Let us consider the additional work in the above algorithm compared to the standard Arnoldi procedure for RRGMRES, where the work involved in j iterations is $O(j^2 n)$ flops. In each iteration, the additional work is dominated by:

1. orthonormalization of the columns of $\widetilde{V}_j$ to v_{j+1} requiring $2pn$ flops,
2. computation of the new F_j requiring $2p^2 n$ flops (assuming AW_p is stored),
3. application of an orthogonal transformation that involves the bottom right $(p+1)\times p$ submatrix requiring about $2p^3$ flops.

Hence, the additional work involved in j iterations is about $2jp(p + 1)n$ flops.

4. Numerical examples. The purpose of this section is to illustrate the performance of the above algorithms with several examples. In all examples we first generate a noise-free system of the form $Ax_{\text{exact}} = b_{\text{exact}}$, and then we add noise to the right-hand side $b = b_{\text{exact}} + e$, where e is a random vector of Gaussian white noise scaled such that $\|e\|_2/\|b_{\text{exact}}\|_2 = \eta$ (where we specify η). For each example we report the relative error $\|x_{\text{exact}} - x^{(j)}\|_2/\|x_{\text{exact}}\|_2$ and the relative residual norm $\|b - Ax^{(j)}\|_2/\|b\|_2$. We use MATLAB and compare combinations of the following algorithms.

- CGLS is the implementation from REGULARIZATION TOOLS [10].
- PCGLS is the subspace-preconditioned CGLS algorithm from REGULARIZATION TOOLS with L an approximation to the second derivative operator.
- RRGMRES is the implementation from REGULARIZATION TOOLS.
- ARRGMRES is our implementation of Augmented RRGMRES from [2].
- R^{3}GMRES is our implementation of Algorithm 2 from Section 3.

All five methods exhibit semi-convergence, but for some of the examples, the slower methods do not reach the minimum error within the number of iterations shown in the plots.

4.1. The solution has a very large component in the augmentation subspace. This example, which was also used in [1, 2], is the test problem `deriv2(n,2)` from REGULARIZATION TOOLS [10] with $n = 32$ and relative noise level $\eta = 10^{-5}$. The augmentation matrix W_2 represents the subspace spanned by the constant and the linear function

$$\mathcal{W}_2 = \text{span}\{w_1, w_2\}, \quad w_1 = (1, 1, \ldots, 1)^\top, \quad w_2 = (1, 2, \ldots, n)^\top.$$

ETNA
Kent State University
http://etna.math.kent.edu

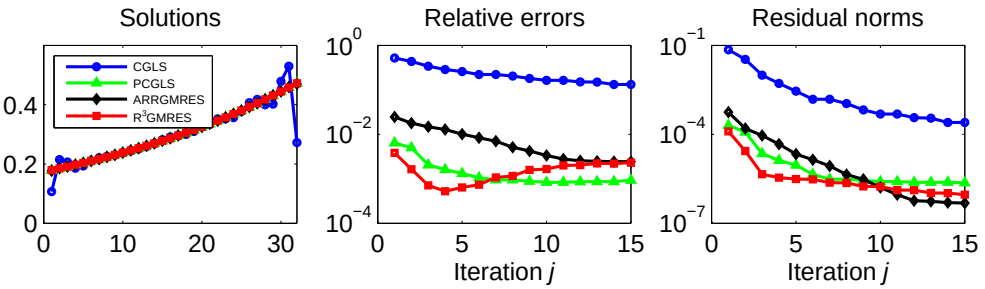


FIG. 4.1. *Example 4.1. Left: best solutions within 15 iterations. Middle and right: convergence histories.*

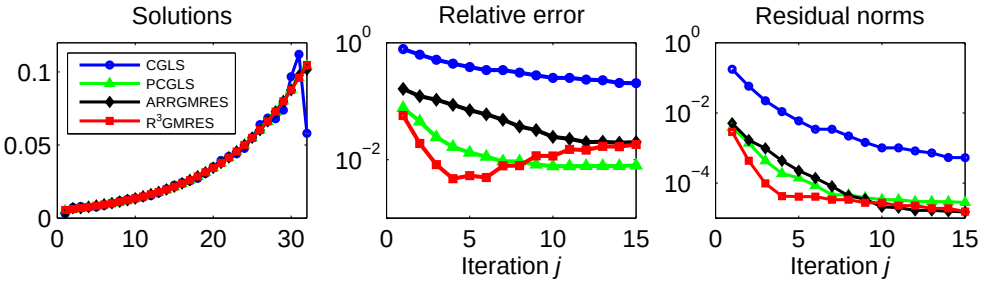


FIG. 4.2. *Example 4.2. Left: best solutions within 15 iterations. Middle and right: convergence histories.*

Here $\|W_2W_2^\top x_{\text{exact}}\|_2/\|x_{\text{exact}}\|_2 = 0.99$ and $\|(I - W_2W_2^\top)x_{\text{exact}}\|_2/\|x_{\text{exact}}\|_2 = 0.035$, so the exact solution x_{exact} has a very large component in $\mathcal{W}_2$. Hence, any iterative regularization method only needs to spend its effort in capturing the small component in $\mathcal{W}_2^\perp$. The results are shown in Figure 4.1.

All right singular vectors of A (not shown here) tend towards zero at both endpoints—this is due to the particular discretization of the problem which assumes zero boundary conditions. Hence, the SVD basis is not well suited for this particular problem whose solution is nonzero at both endpoints. This explains the bad performance of CGLS, which produces filtered SVD solutions as can be seen in Figure 4.1. PCGLS performs much better because $\mathcal{W}_2$ is identical to the null space of L, and therefore any component of the solution in this subspace is immediately captured independent of the iterations.

The residual norms for PCGLS, ARRGMRES, and R^{3}GMRES approach approximately the same level determined by the noise, and all three methods provide regularized solutions with good accuracy; our algorithm R^{3}GMRES has the fastest semi-convergence.

4.2. Fix or improve boundary conditions. In this example, the matrix A is the same as above, but we modified the exact solution `x` from `deriv2(n,2)` by means of `x = x.^3`, which gives a new exact solution with a large first derivative at the right endpoint. The relative noise level here is $\eta = 10^{-4}$, and we use the same augmentation subspace as above. The results are shown in Figure 4.2.

Again the SVD basis is not well suited for this problem, and CGLS produces bad results. The other three iterative methods work well; this time because our choice of W_2 is able to compensate for the "incorrect" or "incompatible" boundary conditions by allowing the regularized solutions to have nonzero values and nonzero derivatives at the endpoints. The error histories of these methods resemble those of the previous example, and again our algorithm R^{3}GMRES has the fastest semi-convergence.

ETNA
Kent State University
http://etna.math.kent.edu

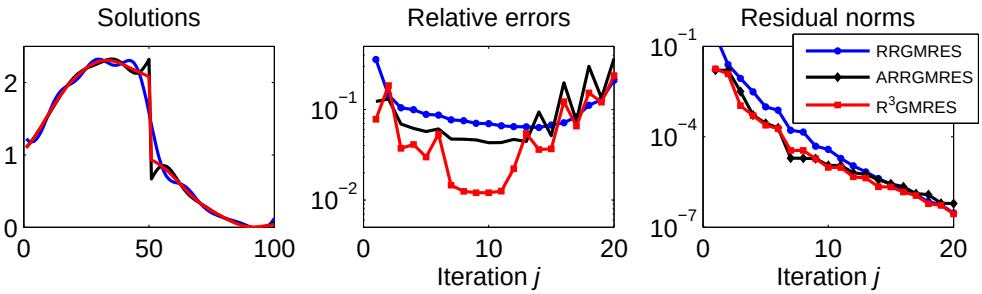


FIG. 4.3. *Example 4.3. Left: best solutions within 20 iterations. Middle and right: error histories.*

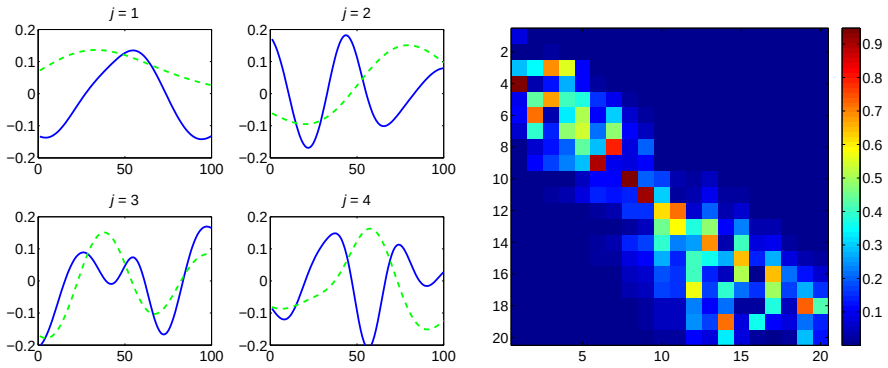


FIG. 4.4. *Example 4.3. Left: the first four orthonormal basis vectors $\bar{v}_{p+j}$ of ARRGMRES (solid blue lines) and v_j of R^{3}GMRES (dashed green lines). Right: an image plot of $|V_{20}^\top \bar{V}_{p+1:p+20}|$, which shows the size of the coefficients of the columns in $\bar{V}_{p+1:p+20}$ in the basis of V_{20}.*

4.3. Capture a single discontinuity. In this example, the matrix A is from the test problem `gravity(n)` in REGULARIZATION TOOLS with $n = 100$, but the exact solution is changed to include a single discontinuity between elements $\ell = 50$ and $\ell + 1 = 51$. We use the matrix W_2 from (2.2), which allows us to represent this discontinuity. The relative noise level is $\eta = 10^{-3}$.

The results are shown in Figure 4.3 for the three methods RRGMRES, ARRGMRES, and R^{3}GMRES, and we see that for all three methods the residual norms decrease monotonically. Clearly, RRGMRES is not well suited for representing the discontinuity because this method does not use the augmentation subspace $\mathcal{W}_p$. Surprisingly ARRGMRES, in spite of the fact that it uses W_2, does not give much better results—all its iterates have certain "ringing" artifacts near the discontinuity. After some iterations, R^{3}GMRES is able to produce much better regularized solutions. We note that the relative error history for R^{3}GMRES is not smooth. This is not an error since it is only the residual norm that has guaranteed monotonic behavior.

To explain why R^{3}GMRES gives better reconstructions than ARRGMRES in this example, we study the basis vectors that, in addition to the columns of W_p, are used to represent the solution. For ARRGMRES, these vectors are the columns $\bar{v}_{p+1}, \bar{v}_{p+2}, \ldots, \bar{v}_{p+j}$ of the matrix $\bar{V}_{p+1:p+j}$, which is an orthonormal basis of $\mathcal{K}_j(P_{V_p}^\perp A, P_{V_p}^\perp Ab)$. For R^{3}GMRES, these vectors are the columns $v_1, v_2, \ldots, v_j$ of V_j, which are orthonormal basis vectors of $\mathcal{K}_j(A, Ab)$.

ETNA
Kent State University
http://etna.math.kent.edu

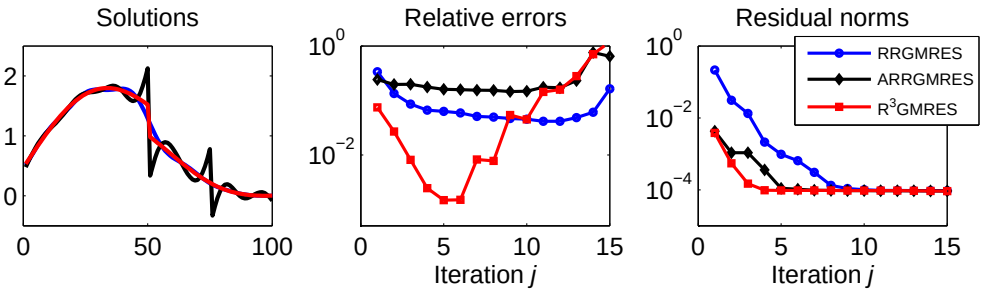


FIG. 4.5. *Example 4.4. Left: best solution obtained within 15 iterations. Middle and right: error histories.*

The left part of Figure 4.4 displays the first four vectors of each method, and we see that some of the very smooth components present in the latter basis are missing from the former. This is confirmed by the image plot in the right part of Figure 4.4, which displays the coefficients of $\bar{v}_{p+1}, \bar{v}_{p+2}, \ldots, \bar{v}_{p+20}$ when expressed in terms of the orthonormal basis $v_1, v_2, \ldots, v_{20}$. This plot clearly shows that the first two smooth components in R^{3}GMRES represented by v_1 and v_2 are missing from the basis in ARRGMRES. Hence, the latter algorithm has difficulties in capturing the smooth components of the solution.

4.4. Handling an additional discontinuity in the augmentation subspace. In this example, we use the same test problem as before, and our goal is to demonstrate what happens if we include a discontinuity in $\mathcal{W}_p$ that is not present in the exact solution. This corresponds to a situation where our prior information tells us about *potential* discontinuities, but not all of them may be present in the given problem. Specifically, in this example there is one discontinuity in the solution but two in $\mathcal{W}_p$. We use an augmentation matrix W_3 similar to that in (2.2) allowing discontinuities between elements 50–51 and 75–76. The noise level is $\eta = 10^{-4}$.

The results are shown in Figure 4.5. RRGMRES does not capture the discontinuity and instead produces smooth reconstructions. ARRGMRES also produces bad solutions; the main reason being that it enforces both discontinuities—both the desired and the undesired. R^{3}GMRES is the only method that introduces a single discontinuity where needed. The explanation is the same as before, namely, that the basis for ARRGMRES lacks the smooth components that are present in the basis for R^{3}GMRES.

5. Conclusions. Inspired by Lothar Reichel's work, we developed an iterative regularization algorithm R^{3}GMRES based on RRGMRES that is able to incorporate prior information in the form of a low-dimension subspace in which the solution is expected to have a large component. Our algorithm computes regularized solutions in a subspace that is the direct sum of this subspace and the Krylov subspace associated with RRGMRES. Numerical examples show that our method gives regularized solutions that are at least as accurate as those computed by other methods, and in most cases our algorithm is faster or more accurate (or both).

Acknowledgements. Our interest in augmented RRGMRES was initiated by discussions with Dr. Nao Kuroiwa, who visited DTU in 2011–2012. Ideas that led to the current algorithm arose in a student project done in collaboration with Emil Brandt Kærgaard.

ETNA
Kent State University
http://etna.math.kent.edu

REFERENCES

[1] J. BAGLAMA AND L. REICHEL, *Decomposition methods for large linear discrete ill-posed problems*, J. Comp. Appl. Math., 198 (2007), pp. 332–343.
[2] ———, *Augmented GMRES-type methods*, Numer. Linear Algebra Appl., 14 (2007), pp. 337–350.
[3] J. BAGLAMA, L. REICHEL, AND D. RICHMOND, *An augmented LSQR method*, Numer. Algorithms, 64 (2013), pp. 263–293.
[4] D. CALVETTI, B. LEWIS, AND L. REICHEL, *GMRES-type methods for inconsistens systems*, Linear Algebra Appl., 316 (2000), pp. 157–169.
[5] ———, *On the regularizing properties of the GMRES method*, Numer. Math, 91 (2002), pp. 605–625.
[6] D. CALVETTI, L. REICHEL, AND A. SHUIBI, *Enriched Krylov subspace methods for ill-posed problems*, Linear Algebra Appl., 362 (2003), pp. 257–273.
[7] ———, *Invertible smoothing preconditioners for linear discrete ill-posed problems*, Appl. Numer. Math., 54 (2005), pp. 135–149.
[8] M. HANKE, *Conjugate Gradient Type Methods for Ill-Posed Problems*, Pitman Research Notes in Mathematics 327, Longman, Harlow, 1995.
[9] M. HANKE AND P. C. HANSEN, *Regularization methods for large-scale problems*, Surveys Math. Industry, 3 (1993), pp. 253-315.
[10] P. C. HANSEN, *Regularization Tools version 4.0 for Matlab 7.3*, Numer. Algorithms, 46 (2007), pp. 189–194.
[11] ———, *Discrete Inverse Problems: Insight and Algorithms*, SIAM, Philadelphia, 2010.
[12] ———, *Oblique projections and standard-form transformations for discrete inverse problems*, Numer. Linear Algebra Appl., 20 (2013), pp. 250–258.
[13] P. C. HANSEN AND T. K. JENSEN, *Smoothing-norm preconditioning for regularizing minimum-residual methods*, SIAM J. Matrix Anal. Appl., 29 (2006/07), pp. 1–14.
[14] M. HOCHSTENBACH AND L. REICHEL, *An iterative method for Tikhonov regularization with a general linear regularization operator*, J. Integral Equations Appl., 22 (2010), pp. 465–482.
[15] T. K. JENSEN AND P. C. HANSEN, *Iterative regularization with minimum-residual methods*, BIT, 47 (2007), pp. 103–120.
[16] M. E. KILMER AND G. W. STEWART, *Iterative Regularization and MINRES*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 613–628.
[17] N. KUROIWA AND T. NODERA, *The adaptive augmented GMRES method for solving ill-posed problems*, ANZIAM J., 50 (2008), pp. C654–C667.
[18] A. NEUMAN, L. REICHEL, AND H. SADOK, *Implementations of range restrited iterative methods for linear discrete ill-posed problems*, Linear Algebra Appl., 436 (2012), pp. 3974–3990.
[19] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, 2003.
[20] A. VAN DER SLUIS AND H. A. VAN DER VORST, *SIRT- and CG-type methods for the iterative solution of sparse linear least-squares problems*, Linear Algebra Appl., 130 (1990), pp. 257–302.