Electronic Transactions on Numerical Analysis.
Volume 55, pp. 532–546, 2022.
Copyright © 2022, Kent State University.
ISSN 1068–9613.
DOI: 10.1553/etna_vol55s532

ETNA
Kent State University and
Johann Radon Institute (RICAM)

A NOTE ON AUGMENTED UNPROJECTED KRYLOV SUBSPACE METHODS*

KIRK M. SOODHALTER†

Abstract. Subspace recycling iterative methods and other subspace augmentation schemes are a successful extension to Krylov subspace methods in which a Krylov subspace is augmented with a fixed subspace spanned by vectors deemed to be helpful in accelerating convergence or conveying knowledge of the solution. Recently, a survey was published, in which a framework describing the vast majority of such methods was proposed [Soodhalter et al., GAMM-Mitt., 43 (2020), Art. e202000016]. In many of these methods, the Krylov subspace is one generated by the system matrix composed with a projector that depends on the augmentation space. However, it is not a requirement that a projected Krylov subspace be used. There are augmentation methods built on using Krylov subspaces generated by the original system matrix, and these methods also fit into the general framework. In this note, we observe that one gains implementation benefits by considering such augmentation methods with unprojected Krylov subspaces in the general framework. We demonstrate this by applying the idea to the R^{3}GMRES method proposed in [Dong et al., Electron. Trans. Numer. Anal., 42 (2014), pp. 136–146] to obtain a simplified implementation and to connect that algorithm to early augmentation schemes based on flexible preconditioning [Saad, SIAM J. Matrix Anal. Appl., 18 (1997)].

Key words. Krylov subspaces, augmentation, recycling, discrete ill-posed problems

AMS subject classifications. 65F10, 65F50, 65F08

1. Introduction. Augmented and recycled Krylov subspace methods have been proposed for accelerating iterative methods for solving a linear system (e.g., [18]) or a sequences of linear systems (see, e.g., [21]) by approximating the solution to each linear system from the sum of a Krylov subspace $\mathcal{V}_j$ and a fixed subspace $\mathcal{U}$. The survey [29] details many instances of such methods in the literature and proposes a framework which describes their general mechanics and common mathematical structure they all share. In most cases, the Krylov subspace used by such a method is a *projected* Krylov subspace, meaning the matrix is composed with a projector which depends on $\mathcal{U}$. However, there are examples in the literature of augmented methods which use an unprojected Krylov subspace built using only the matrix; see, e.g., [7, 9]. Such methods necessarily also fit into the framework but are not generally described as such. In this note, we focus on one such method, R^{3}GMRES, proposed in [7]; we show how considering it as an augmented method in the framework from [29] allows for a simpler implementation built on well-understood algorithmic blocks from classical GMRES [27]. In addition, we point out that R^{3}GMRES can be related to an older augmentation scheme built on flexible preconditioning [26].

2. Background. In principle, the R^{3}GMRES algorithm can be applied to any square, discrete linear problem, but it is proposed specifically to treat discrete ill-posed problems. Therefore, we begin with a brief description of the ill-posed problem setting.

Ill-posed problems arise often in the context of scientific applications in which one cannot directly observe the object or quantity of interest. However, indirect observations or measurements can be made. We restrict ourselves to the linear case, whereby the unobservable quantity of interest and the measured data can be related by a linear operator. In this note, we consider a discretized, finite-dimensional version of this problem,

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \text{with} \quad \mathbf{A} \in \mathbb{R}^{n\times n} \quad \text{and} \quad \mathbf{b} \in \mathbb{R}^{n}. \tag{2.1}$$

*Received June 23, 2021. Accepted March 3, 2022. Published online on June 4, 2022. Recommended by Silvia Gazzola.

†School of Mathematics, Trinity College Dublin, The University of Dublin, College Green, Dublin 2, Ireland (ksoodha@maths.tcd.ie).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

The vector $\mathbf{b}$ represents the observed data, obtained from measurements (uncontaminated by measurement noise), and $\mathbf{x}$ represents the quantity of interest, which cannot be directly observed. The matrix $\mathbf{A}$ is generally taken to have large condition number and singular values that decrease smoothly with no breaks to indicate a separation between the well-posed and ill-posed parts of the discrete problem. As this is a finite-dimensional discretized problem, we expect perturbations in the right-hand side to produce bounded perturbations in the reconstructed solution that are generally large enough to render the reconstructed solution useless. Thus, we must consider regularization methods. In this note, we concern ourselves with some GMRES-based regularization techniques for sparse, large-scale problems, but there is an extensive literature on the topics of Krylov subspace methods and hybrid methods; see, e.g., the surveys [3, 12].

In the next section, we review some general mathematics behind Krylov subspace methods. We explain briefly GMRES before turning our attention to augmented Krylov subspace methods. In Section 3.4, we review the basic mechanics of augmented/recycled Krylov subspace methods, particularly in the context of the framework proposed in [29]. In Section 4, we show how the R^{3}GMRES method can be simplified by casting it in this framework. Finally, we demonstrate the behavior of the new implementation with some numerical experiments in Section 5.

NOTATION 1. In this paper, we denote by $\mathbf{I}_\ell \in \mathbb{R}^{\ell\times\ell}$ the identity matrix acting on $\mathbb{R}^\ell$, and if the dimensions are understood from context, we simply write $\mathbf{I}$. Additionally, $\underline{\mathbf{I}}_\ell \in \mathbb{R}^{(\ell+1)\times\ell}$ denotes the same identity matrix but with an extra row of zeros appended at the bottom. The vector $\mathbf{x}_0 \in \mathbb{R}^n$ denotes the initial approximation. We denote the initial error $\boldsymbol{\eta}_0 = \mathbf{x} - \mathbf{x}_0$ and the initial residual $\mathbf{r}_0 = \mathbf{A}\boldsymbol{\eta}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$. The vector $\mathbf{e}_i$ denotes the ith canonical basis vectors, whose length is defined by context.

3. Background on Krylov subspace methods.

In this section, we begin with a general description of Krylov subspace iterative methods, specifically the Generalized Minimum Residual Method (GMRES). We then offer a brief review of augmented Krylov methods, which have been developed both in the well-posed and ill-posed problems literature. We observe that there has been some overlap in the developments in the two communities.

3.1. Krylov Subspace Methods.

Krylov subspace iterative methods are a well-known class of methods for the solution of linear systems as well as other types of problems. For solving a linear system of the form (2.1) with $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{b} \in \mathbb{R}^n$, one builds the Krylov subspace

$$\mathcal{K}_j(\mathbf{A}, \mathbf{r}_0) = \left\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \mathbf{A}^2\mathbf{r}_0, \ldots, \mathbf{A}^{j-1}\mathbf{r}_0\right\}$$

iteratively (at the cost of one matrix-vector product per iteration). At iteration j, a correction $\mathbf{t}_j \in \mathcal{K}_j(\mathbf{A}, \mathbf{r}_0)$ is selected according to some constraints on the residual $\mathbf{r}_j = \mathbf{b} - \mathbf{A}\mathbf{x}_j$, where $\mathbf{x}_j = \mathbf{x}_0 + \mathbf{t}_j$ is the jth approximation. We call $\mathbf{t}_j$ a *correction* and the space from which it is drawn the *correction space*. In this paper, we focus on the Generalized Minimum Residual Method (GMRES) [27] in which we select

$$\mathbf{t}_j \in \mathcal{K}_j(\mathbf{A}, \mathbf{r}_0) \quad \text{such that} \quad \mathbf{r}_j \perp \mathbf{A}\mathcal{K}_j(\mathbf{A}, \mathbf{r}_0). \tag{3.1}$$

Such an orthogonality condition for the residual is called a *Petrov-Galerkin condition*. Methods with such a residual orthogonality constraint are often called *residual projection methods* because the constraint leads to a projection (oblique or orthogonal) of the residual. This particular constraint is equivalent to solving the residual minimization problem

$$\mathbf{t}_j = \underset{\mathbf{t}\in\mathcal{K}_j(\mathbf{A},\mathbf{r}_0)}{\operatorname{argmin}} \left\|\mathbf{b} - \mathbf{A}(\mathbf{x}_0 + \mathbf{t})\right\|_2, \tag{3.2}$$

and the residual is projected orthogonally (implicitly) to obtain the updated approximation, with

$$\mathbf{t}_j = \mathbf{P}_{\mathcal{K}_j}\boldsymbol{\eta}_0 \quad \text{and} \quad \mathbf{r}_j = \left(\mathbf{I} - \mathbf{Q}_{\mathcal{K}_j}\right)\mathbf{r}_0,$$

with $\mathbf{P}_{\mathcal{K}_j}$ being the $\left(\mathbf{A}^T\mathbf{A}\right)$-orthogonal projector onto $\mathcal{K}_j\left(\mathbf{A}, \mathbf{r}_0\right)$ and $\mathbf{Q}_{\mathcal{K}_j}$ being the orthogonal projector onto $\mathbf{A}\mathcal{K}_j\left(\mathbf{A}, \mathbf{r}_0\right)$. During the iteration, one builds an orthonormal basis for $\mathcal{K}_j(\mathbf{A}, \mathbf{r}_0)$ one vector at a time using the Arnoldi process. At iteration j, the process has generated

$$\mathbf{V}_{j+1} = \begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_{j+1}\end{bmatrix} \in \mathbb{R}^{n\times(j+1)} \quad \text{and} \quad \underline{\mathbf{H}}_j \in \mathbb{R}^{(j+1)\times j},$$

where the columns of $\mathbf{V}_j$ form an orthonormal basis for $\mathcal{K}_j(\mathbf{A}, \mathbf{r}_0)$ and $\underline{\mathbf{H}}_j$ is an upper Hessenberg matrix (with zeros below the first subdiagonal) containing the orthogonalization coefficients. From the construction of the basis, we get the Arnoldi relation

$$\mathbf{A}\mathbf{V}_j = \mathbf{V}_{j+1}\underline{\mathbf{H}}_j = \mathbf{V}_j\mathbf{H}_j + h_{j+1,j}\mathbf{v}_{j+1}\mathbf{e}_j^T, \tag{3.3}$$

where $\mathbf{H}_j \in \mathbb{R}^{j\times j}$ is simply the first j rows of $\underline{\mathbf{H}}_j$. From (3.3), one can reduce the minimization (3.2) to a smaller $(j+1)\times j$ least-squares minimization problem

$$\mathbf{y}_j = \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^j} \left\| \underline{\mathbf{H}}_j\mathbf{y} - \beta\mathbf{e}_1^{(j+1)} \right\|_2 \quad \text{and} \quad \mathbf{t}_j = \mathbf{V}_j\mathbf{y}_j, \tag{3.4}$$

where $\beta = \|\mathbf{r}_0\|_2$. The standard implementation dictates that we compute the QR-factorization $\underline{\mathbf{H}}_j = \mathbf{Q}_j\underline{\mathbf{R}}_j$ using Givens rotations, where $\mathbf{Q}_j \in \mathbb{R}^{(j+1)\times(j+1)}$ is an orthogonal matrix and $\underline{\mathbf{R}}_j \in \mathbb{R}^{(j+1)\times j}$ is upper triangular. Via the economy QR-factorization of $\underline{\mathbf{H}}_j$, we can recast the minimization in (3.4) as the solution of an upper-triangular linear system

$$\mathbf{R}_j\mathbf{y}_j = \left(\mathbf{Q}_j^T(\beta\mathbf{e}_1)\right)_{1:j},$$

where $\mathbf{R}_j \in \mathbb{R}^{j\times j}$ is simply the first j rows of $\underline{\mathbf{R}}_j$ and $(\cdot)_{1:j}$ denotes taking the first j rows of the argument. This can be used to develop a progressive formulation of GMRES, but more importantly, it allows one to monitor the residual norm without computing the GMRES approximation at each iteration. One can show that the jth residual norm is simply the $(j+1)$st row of $\mathbf{Q}_j^T(\beta\mathbf{e}_1)$ [27].

3.2. Range-restricted Krylov subspace methods. In the context of ill-posed problems, range-restricted methods have been proposed, wherein the Krylov subspace used is $\mathcal{K}_j(\mathbf{A}, \mathbf{A}\mathbf{r}_0)$ rather than simply generating it with the residual $\mathbf{r}_0$. The rationale in this setting is that the right-hand side (and therefore the initial residual) may be profoundly noise-polluted in such a way that reduces the effectiveness of the Krylov subspace method. In such problems, the matrix $\mathbf{A}$ is a discretized version of an operator that often has smoothing properties, meaning $\mathbf{A}\mathbf{r}_0$ is a smoothed version of the initial data, and using the range-restricted subspace will produce a more stable iteration. Range-restricted versions of GMRES [19, 20, 25] and MINRES [8] have been proposed, with the latter being a practical realization of the MR2 method discussed in Hanke's monograph [15].

3.3. Augmented methods for well- and ill-posed problems. Augmented Krylov subspace methods have been discussed in both the well- and ill-posed problems communities, though in each with different goals in mind. The term *augmented Krylov subspace method* describes here an iterative method in which, in addition to generating a Krylov subspace, one wishes to include vectors in the correction space deemed useful for either accelerating the convergence to the solution or improving the quality of the approximation delivered by the method.

For well-posed problems, these vectors may span a subspace which has been determined to have strongly contributed to the speed of convergence [5] or to attempt to damp the influence of certain parts of the spectrum of the operator [18, 21]. For ill-posed problems, this strategy has also been shown to be effective in the case that, e.g., the noise level is rather low, as the solution may require many iterations [17].

However, in the context of large-scale, discrete ill-posed problems, one may also augment with vectors representing known features of the image, usually those which are highly local, such as discontinuous jumps or areas of high gradient, which an iterative method based on a Krylov subspace method may have difficulty resolving [1, 2, 7]. Recycling-based strategies have also been shown to be effective for some such applications [17]. Recently, using the framework from [29], augmented methods were analyzed formally [23] as regularization methods.

3.4. Subspace augmentation via a minimization constraint. We briefly present a general residual constraint framework through which the methods in question can be viewed. For a more complete view of this framework, see [29] in terms of residual constraints on top of the existing work in [10, 11, 13, 14].

In this framework, we approach augmented methods by approximating the correction over the sum of two subspaces $\mathcal{U}$, which is *fixed*, and $\mathcal{V}_j$, which is *built iteratively* (i.e., it generally is some sort of Krylov subspace). In this note, we consider the special case that we apply a residual-minimizing constraint. This technique is a straightforward generalization of the minimum residual projection constraint (3.1), i.e., we require

$$\mathbf{b} - \mathbf{A}\left(\mathbf{x}_0 + \mathbf{s}_j + \mathbf{t}_j\right) \perp \mathbf{A}\left(\mathcal{U} + \mathcal{V}_j\right). \tag{3.5}$$

Let $\mathbf{U} \in \mathbb{R}^{n\times k}$ be such that $\mathcal{R}(\mathbf{U}) = \mathcal{U}$. This residual constraint underpins (either implicitly or explicitly) many augmented GMRES-type methods. Associated to this constraint are, respectively, the $\mathbf{A}^T\mathbf{A}$-orthogonal projector onto $\mathcal{U}$ and the orthogonal projector onto $\mathcal{C} := \mathbf{A}\mathcal{U}$

$$\mathbf{\Pi} = \mathbf{U}\left(\mathbf{U}^T\mathbf{A}^T\mathbf{A}\mathbf{U}\right)^{-1}\mathbf{U}^T\mathbf{A}^T\mathbf{A} \quad \text{and} \quad \mathbf{\Phi} = \mathbf{A}\mathbf{U}\left(\mathbf{U}^T\mathbf{A}^T\mathbf{A}\mathbf{U}\right)^{-1}\mathbf{U}^T\mathbf{A}^T.$$

If a method minimizes the residual over a sum of subspaces, then it *necessarily* fits into the augmentation framework, regardless of how $\mathcal{V}_j$ is generated. It is pointed out in [29] that regardless of the choice of $\mathcal{V}_j$, this residual minimization over the sum of subspaces can be reduced and reformulated as selecting $\mathbf{t} \approx \mathbf{t}_j \in \mathcal{V}_j$ to minimize the residual of the projected problem

$$(\mathbf{I} - \mathbf{\Phi})\mathbf{A}\mathbf{t} = (\mathbf{I} - \mathbf{\Phi})\mathbf{r}_0 \tag{3.6}$$

and setting $\mathbf{x}_j = \mathbf{x}_0 + \mathbf{\Pi}\boldsymbol{\eta}_0 + (\mathbf{I} - \mathbf{\Pi})\mathbf{t}_j$, where we note that the action of $\mathbf{\Pi}$ on $\boldsymbol{\eta}_0$ can be computed efficiently without knowing $\boldsymbol{\eta}_0$. We show that this can lead to simplified implementations of such methods, particularly as it relates to methods which augment *unprojected* Krylov subspace methods.

For methods such as GMRES-DR [18] and GCRO-type methods, e.g., [4, 5, 21], the iteratively generated Krylov subspace matches with the projected subproblem (3.6) with $\mathcal{V}_j = \mathcal{K}_j\left(\left(\mathbf{I} - \mathbf{\Phi}\right)\mathbf{A}, \left(\mathbf{I} - \mathbf{\Phi}\right)\mathbf{r}_0\right)$. Augmented methods based on range-restricted GMRES, e.g., [2], use $\mathcal{V}_j = \mathcal{K}_j\left(\left(\mathbf{I} - \mathbf{\Phi}\right)\mathbf{A}, \left(\mathbf{I} - \mathbf{\Phi}\right)\mathbf{A}\mathbf{r}_0\right)$. With GCRO-based augmented (range-restricted) GMRES, one can implement either one small minimization problem over the

ETNA
Kent State University and
Johann Radon Institute (RICAM)

augmented subspace or, by directly using the above framework, approximate the solution of (3.6) by a GMRES minimization followed by a projection, as described above. Let

$$\mathcal{V}_j = \mathcal{K}_j\left(\left(\mathbf{I}-\mathbf{\Phi}\right)\mathbf{A}, \left(\mathbf{I}-\mathbf{\Phi}\right)\mathbf{w}_0\right), \quad \text{where} \quad \mathbf{w}_0 \in \left\{\mathbf{r}_0,\ \mathbf{A}\mathbf{r}_0\right\}.$$

Let $\mathbf{V}_j$ be generated by the Arnoldi process so that we have

$$\left(\mathbf{I}-\mathbf{\Phi}\right)\mathbf{A}\mathbf{V}_j = \mathbf{V}_{j+1}\underline{\mathbf{H}}_j.$$

ASSUMPTION 1. *The matrix* $\mathbf{U}$ *is scaled so that* $\mathbf{C} = \mathbf{A}\mathbf{U}$ *has orthonormal columns, i.e.,* $\mathbf{C}^T\mathbf{C} = \mathbf{I}_k$ *and* $\mathbf{\Phi} = \mathbf{C}\mathbf{C}^T$. *This is not mathematically necessary, but it allows for various algorithmic simplifications.*

In [21], the authors approach recycling by deriving a modified Arnoldi relation

$$\mathbf{A}\begin{bmatrix}\mathbf{U} & \mathbf{V}_j\end{bmatrix} = \begin{bmatrix}\mathbf{C} & \mathbf{V}_{j+1}\end{bmatrix}\underline{\mathbf{G}}_j, \quad \text{where} \quad \underline{\mathbf{G}}_j = \begin{bmatrix}\mathbf{I}_k & \mathbf{B}_j \\ & \underline{\mathbf{H}}_j\end{bmatrix} \quad \text{and} \quad \mathbf{B}_j = \mathbf{C}^T\mathbf{A}\mathbf{V}_j.$$

From this, one can satisfy (3.5) by solving the small least-squares problem

$$(\mathbf{z}_j, \mathbf{y}_j) = \underset{\substack{\mathbf{u}\in\mathbb{R}^k \\ \mathbf{v}\in\mathbb{R}^j}}{\operatorname{argmin}} \left\| \begin{bmatrix}\mathbf{C} & \mathbf{V}_{j+1}\end{bmatrix}^T \mathbf{r}_0 - \underline{\mathbf{G}}_j \begin{bmatrix}\mathbf{z} \\ \mathbf{y}\end{bmatrix} \right\|_2. \tag{3.7}$$

This is in actuality not necessary for implementing the method, since it can be decoupled to solve a GMRES small least-squares problem for $\mathbf{y}_j$ which then enables the solution of $\mathbf{z}_j$ by back substitution; i.e.,

$$\mathbf{y}_j = \underset{\mathbf{y}\in\mathbb{R}^j}{\operatorname{argmin}} \left\|\beta\mathbf{e}_1 - \underline{\mathbf{H}}_j\mathbf{y}\right\| \quad \text{and} \quad \mathbf{z}_j = \mathbf{C}^T\mathbf{r}_0 - \mathbf{B}_j\mathbf{y}_j.$$

However, (3.7) is useful as a comparison to the coupled minimization in the proposed implementation of augmented unprojected (range-restricted) GMRES, which we discuss below.

3.5. Augmenting unprojected Krylov subspaces. In both the well- and ill-posed problems community, augmented methods have been proposed wherein an unprojected Krylov subspace is used, in [26] $\mathcal{V}_j = \mathcal{K}_j\left(\mathbf{A}, \mathbf{r}_0\right)$ and in [7] $\mathcal{V}_j = \mathcal{K}_j\left(\mathbf{A}, \mathbf{A}\mathbf{r}_0\right)$. We discuss briefly some implementation details of these methods *which are relevant to the present note*, but one should read the cited papers and references therein for complete implementation details. It has also been observed that under certain circumstances in which there are strict constraints on the amount of computations one can perform per iteration, an unprojected augmented method may be preferred (or indeed be the only option); see [24], which builds on [9].

3.5.1. Flexible GMRES-based augmentation for well-posed problems. In [26], Saad proposes augmenting an already constructed Krylov subspace $\mathcal{K}_j(\mathbf{A}, \mathbf{r}_0)$ with a subspace $\mathcal{W} = \operatorname{span}\left\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k\right\}$ by treating the basis vectors of $\mathcal{W}$ as those resulting from the action of successive implicit flexible preconditioners. The augmentation process is embedded in an iteration of flexible GMRES. This minimum residual method can be described in the language of the framework by identifying that the correction space in this setting is $\underbrace{\mathcal{K}_j(\mathbf{A}, \mathbf{r}_0)}_{\mathcal{V}_j} + \underbrace{\mathcal{W}}_{\mathcal{U}}$ and the constraint space is $\underbrace{\mathbf{A}\mathcal{K}_j(\mathbf{A}, \mathbf{r}_0)}_{\widetilde{\mathcal{V}}_j} + \underbrace{\mathbf{A}\mathcal{W}}_{\widetilde{\mathcal{U}}}$, where the flexible Arnoldi process produces an orthonormal basis for the constraint space. An outline of this method is shown in Algorithm 1.

Algorithm 1: One cycle of Flexible GMRES-based augmentation from [26].

Input : $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b}, \mathbf{x}_0 \in \mathbb{R}^n$, $\mathbf{W} \in \mathbb{R}^{n\times k}$, $m > 0$

$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$; $\beta = \|\mathbf{r}_0\|$; $\mathbf{v}_1 = \mathbf{r}_0/\beta$;

for $i = 1, 2, \ldots, m + k$ **do**

  if $i < m$ **then**

    $\mathbf{v}_{i+1} = \mathbf{A}\mathbf{v}_i$;

  else

    $\mathbf{v}_{i+1} = \mathbf{A}\mathbf{w}_{i-m+1}$;

  end

  for $j = 1, 2, \ldots i$ **do**

    $h_{ji} = \mathbf{v}_j^T \mathbf{v}_{i+1}$;

    $\mathbf{v}_{i+1} \leftarrow \mathbf{v}_{i+1} - h_{ji}\mathbf{v}_j$;

  end

  $h_{i+1,i} = \|\mathbf{v}_{i+1}\|$;

  $\mathbf{v}_{i+1} \leftarrow \mathbf{v}_{i+1}/h_{i+1,i}$;

  $\mathbf{y} = \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^i} \|\beta\mathbf{e}_1 - \underline{\mathbf{H}}_i\mathbf{y}\|$;

end

$\mathbf{x} = \mathbf{x}_0 + \mathbf{V}_m\mathbf{y}(\mathtt{1:m}) + \mathbf{W}\mathbf{y}(\mathtt{m+1:m+k})$;

3.5.2. Augmentation of (range-restricted) methods for solving ill-posed problems. In the context of solving discrete ill-posed problems using augmented iterative techniques, it has been asserted in [7] that it may be preferable to employ augmentation techniques with an unprojected Krylov subspace. This is in part motivated by the use of projected Krylov subspaces in [1]. The authors argue that the subspace $\mathcal{U}$ should contain (approximations of) known features of the image. However, if these features are poor approximations of image features (e.g., a misplaced discontinuity), then it is asserted that the use of a projected Krylov subspace can cause the iteration to semi-converge to a poor quality solution. Conversely, for solving a well-posed problem, the iteration would eventually recover and converge. The authors suggest using $\mathcal{V}_j = \mathcal{K}_j(\mathbf{A}, \mathbf{w}_0)$, with $\mathbf{w}_0 \in \{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0\}$ preferring $\mathbf{w}_0 = \mathbf{A}\mathbf{r}_0$ (i.e., a range-restricted method), as it tends to yield superior performance for their experiments [7].

REMARK 3.1. For the case $\mathbf{w}_0 = \mathbf{r}_0$, we observe that the method shown in [7] is mathematically equivalent to the augmentation in a flexible preconditioning framework proposed by Saad [26], an equivalence noted in [29].

4. Framework perspective allows for a simplified unprojected augmented GMRES. Again, for implementation purposes, we invoke Assumption 1, i.e., that $\mathbf{C} = \mathbf{A}\mathbf{U}$ has orthonormal columns. At each iteration of the Arnoldi process for $\mathcal{K}_j(\mathbf{A}, \mathbf{w}_0)$, the method proposed in [7] requires an orthonormal basis for the columns of $\mathbf{A}\begin{bmatrix}\mathbf{V}_j & \mathbf{U}\end{bmatrix}$. Unlike with GCRO-type methods, this does not come for free since the Krylov subspace is unprojected. We show in the following subsection that approaching this method from the framework point of view allows us to avoid the algorithmic complication of this orthogonalization. The framework enables us to solve for least-squares approximate solutions of the projected problem (3.6) over the unprojected Krylov subspace $\mathcal{K}_j(\mathbf{A}, \mathbf{w}_0)$ and then obtain an additional correction over $\mathcal{U}$ to obtain the full approximation without additional orthogonalization complications. Furthermore, this new formulation allows for the estimation of the residual norm, meaning that similar to an efficient implementation of GMRES, neither the full approximation nor the residual need to be computed until possible convergence has been detected.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

We derive a simplified version of R[3]GMRES in [7]. We begin our derivation similar to that in [7] by assuming that one must progressively orthogonalize $\mathbf{C}$ against the Arnoldi vectors, but through our derivation we show that this is actually not necessary.

Let

$$\widehat{\mathbf{C}}_1 = \mathbf{C} - \mathbf{v}_1 \left(\mathbf{v}_1^T \mathbf{C}\right) \quad \text{and} \quad \widehat{\mathbf{C}}_1 = \mathbf{C}_1 \mathbf{F}_1 \text{ (skinny } \mathtt{QR}\text{-factorization).}$$

At each iteration i, this orthogonalization must be updated after $\mathbf{v}_{i+1}$ has been generated, and this can be performed recursively

$$\widehat{\mathbf{C}}_{i+1} = \mathbf{C}_i - \mathbf{v}_{i+1} \left(\mathbf{v}_{i+1}^T \mathbf{C}_i\right) \quad \text{and} \quad \widehat{\mathbf{C}}_{i+1} = \mathbf{C}_{i+1} \mathbf{F}_{i+1} \text{ (skinny } \mathtt{QR}\text{-factorization).}$$

From this, one gets the new modified Arnoldi factorization

$$\mathbf{A} \begin{bmatrix} \mathbf{V}_j & \mathbf{U} \end{bmatrix} = \begin{bmatrix} \mathbf{V}_{j+1} & \mathbf{C}_j \end{bmatrix} \widehat{\mathbf{G}}_j \quad \text{with} \quad \widehat{\mathbf{G}}_j = \begin{bmatrix} \underline{\mathbf{H}}_j & \mathbf{D}_j \\ & \mathbf{F}_j \end{bmatrix},$$

where $\mathbf{D}_j = \mathbf{V}_{j+1}^T \mathbf{C}$ and $\mathbf{F}_j = \mathbf{C}_j^T \mathbf{C}$. One observes that $\mathbf{D}_j$ can be constructed iteratively, as

$$\mathbf{D}_j = \mathbf{V}_{j+1}^T \mathbf{C} = \begin{bmatrix} \mathbf{V}_j^T \mathbf{C} \\ \mathbf{v}_{j+1}^T \mathbf{C} \end{bmatrix} = \begin{bmatrix} \mathbf{D}_{j-1} \\ \mathbf{d}_j \end{bmatrix},$$

where $\mathbf{d}_j = \mathbf{v}_{j+1}^T \mathbf{C}$.

As with GCRO-based methods, the minimization constraint (3.5) reduces to a small least-squares problem similar to (3.7), namely

$$(\mathbf{z}_j, \mathbf{y}_j) = \operatorname*{argmin}_{\substack{\mathbf{z} \in \mathbb{R}^k \\ \mathbf{y} \in \mathbb{R}^j}} \left\| \begin{bmatrix} \mathbf{V}_{j+1} & \mathbf{C}_j \end{bmatrix}^T \mathbf{r}_0 - \widehat{\mathbf{G}}_j \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \right\|_2 . \tag{4.1}$$

This is the minimization that is then explicitly solved in [7]. However, just like the GCRO-based methods, this minimization over the sum of subspaces can be rewritten as the approximation of the solution of a projected subproblem (3.6) over $\mathcal{V}_j$ whose solution is then projected onto $\mathcal{U}$ to get $\mathbf{s}_j$. The difference here is that $\mathcal{V}_j$ is the Krylov subspace associated to the unprojected problem; i.e., $\mathcal{V}_j = \mathcal{K}_j(\mathbf{A}, \mathbf{w}_0)$.

The method proposed in [7] is a residual minimization over the sum of two spaces; thus it must fit into the framework introduced in Section 3.4. Our task is to understand how this residual minimization over the sum of two spaces can be rewritten as a minimization for a projected subproblem, just as we have described for GCRO-based methods. This brings us to the main result.

THEOREM 4.1. *Let $\mathcal{V}_j = \mathcal{K}_j(\mathbf{A}, \mathbf{w}_0)$, where $\mathbf{w}_0 \in \{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0\}$. Minimizing the residual over the sum of spaces $\mathcal{U} + \mathcal{V}_j$ as described in both [26] and [7] is equivalent to computing $\mathbf{y}_j$ satisfying*

$$\left(\underline{\mathbf{H}}_j^T \underline{\mathbf{H}}_j - \underline{\mathbf{H}}_j^T \mathbf{D}_j \mathbf{D}_j^T \underline{\mathbf{H}}_j\right) \mathbf{y}_j = \underline{\mathbf{H}}_j^T \left(\mathbf{V}_{j+1}^T (\mathbf{I} - \boldsymbol{\Phi})\, \mathbf{r}_0\right) \tag{4.2}$$

and $\mathbf{z}_j = \mathbf{C}^T \mathbf{r}_0 - \mathbf{D}_j^T \underline{\mathbf{H}}_j \mathbf{y}_j$. Furthermore, this is equivalent to finding $\mathbf{t}_j = \mathbf{V}_j \mathbf{y}_j \in \mathcal{V}_j$ which satisfies a least-squares minimization applied to the projected subproblem (3.6), namely

$$\text{select} \quad \mathbf{t}_j \in \mathcal{V}_j \quad \text{such that} \quad \left\| (\mathbf{I} - \boldsymbol{\Phi}) \left(\mathbf{b} - \mathbf{A}\left(\mathbf{x}_0 + \mathbf{t}_j\right)\right) \right\| \text{ is minimized.} \tag{4.3}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Proof. One can take a couple of different approaches to see how one solves the projected subproblem. Here we follow the approach in [22], wherein we form the normal equations of (4.1)

$$\begin{bmatrix}\mathbf{F}_j^T & \mathbf{D}_j^T\\ & \underline{\mathbf{H}}_j^T\end{bmatrix}\begin{bmatrix}\mathbf{F}_j & \\ \mathbf{D}_j & \underline{\mathbf{H}}_j\end{bmatrix}\begin{bmatrix}\mathbf{z}_j\\ \mathbf{y}_j\end{bmatrix} = \begin{bmatrix}\mathbf{F}_j^T & \mathbf{D}_j^T\\ & \underline{\mathbf{H}}_j^T\end{bmatrix}\begin{bmatrix}\mathbf{C}^T\mathbf{r}_0\\ \mathbf{V}_{j+1}^T\mathbf{r}_0\end{bmatrix} \iff$$
$$\begin{bmatrix}\mathbf{F}_j^T\mathbf{F}_j + \mathbf{D}_j^T\mathbf{D}_j & \mathbf{D}_j^T\underline{\mathbf{H}}_j\\ \underline{\mathbf{H}}_j^T\mathbf{D}_j & \underline{\mathbf{H}}_j^T\underline{\mathbf{H}}_j\end{bmatrix}\begin{bmatrix}\mathbf{z}_j\\ \mathbf{y}_j\end{bmatrix} = \begin{bmatrix}\left(\mathbf{D}_j^T\mathbf{V}_{j+1}^T + \mathbf{F}_j^T\mathbf{C}_j^T\right)\mathbf{r}_0\\ \underline{\mathbf{H}}_j^T\mathbf{V}_{j+1}^T\mathbf{r}_0\end{bmatrix}.$$

A block LU-factorization of the system matrix allows us to eliminate $\mathbf{z}_j$ from the second equation yielding the equations

$$\begin{aligned}&\left(\mathbf{F}_j^T\mathbf{F}_j + \mathbf{D}_j^T\mathbf{D}_j\right)\mathbf{z}_j + \mathbf{D}_j^T\underline{\mathbf{H}}_j\mathbf{y}_j = \left(\mathbf{D}_j^T\mathbf{V}_{j+1}^T + \mathbf{F}_j^T\mathbf{C}_j^T\right)\mathbf{r}_0 \quad \text{and}\\ &\left(\underline{\mathbf{H}}_j^T\underline{\mathbf{H}}_j - \underline{\mathbf{H}}_j^T\mathbf{D}_j\left(\mathbf{D}_j^T\mathbf{D}_j + \mathbf{F}_j^T\mathbf{F}_j\right)^{-1}\mathbf{D}_j^T\underline{\mathbf{H}}_j\right)\mathbf{y}_j\\ &\qquad = \underline{\mathbf{H}}_j^T\mathbf{V}_{j+1}^T\mathbf{r}_0 - \underline{\mathbf{H}}_j^T\mathbf{D}_j\left(\mathbf{D}_j^T\mathbf{D}_j + \mathbf{F}_j^T\mathbf{F}_j\right)^{-1}\left(\mathbf{D}_j^T\mathbf{V}_{j+1}^T + \mathbf{F}_j^T\mathbf{C}_j^T\right)\mathbf{r}_0.\end{aligned}$$

Observe now that if we substitute the definitions of $\mathbf{D}_j$ and $\mathbf{F}_j$ into the latter equations, we get that

$$\mathbf{D}_j^T\mathbf{D}_j + \mathbf{F}_j^T\mathbf{F}_j = \mathbf{C}^T\left(\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T + \mathbf{C}_j\mathbf{C}_j^T\right)\mathbf{C}.$$

By design, we have that $\mathcal{R}\left(\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T\mathbf{C}\right) \oplus \mathcal{R}\left(\mathbf{C}_j\right) = \mathcal{C}$, which implies that

$$\left(\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T + \mathbf{C}_j\mathbf{C}_j^T\right)\mathbf{C} = \mathbf{C}, \tag{4.4}$$

and by Assumption 1, we get

$$\begin{aligned}\mathbf{z}_j &= \left(\mathbf{D}_j^T\mathbf{V}_{j+1}^T + \mathbf{F}_j^T\mathbf{C}_j^T\right)\mathbf{r}_0 - \mathbf{D}_j^T\underline{\mathbf{H}}_j\mathbf{y}_j \quad \text{and}\\ \left(\underline{\mathbf{H}}_j^T\underline{\mathbf{H}}_j - \underline{\mathbf{H}}_j^T\mathbf{D}_j\mathbf{D}_j^T\underline{\mathbf{H}}_j\right)\mathbf{y}_j &= \underline{\mathbf{H}}_j^T\mathbf{V}_{j+1}^T\mathbf{r}_0 - \underline{\mathbf{H}}_j^T\mathbf{D}_j\left(\mathbf{D}_j^T\mathbf{V}_{j+1}^T + \mathbf{F}_j^T\mathbf{C}_j^T\right)\mathbf{r}_0.\end{aligned} \tag{4.5}$$

We finish by showing that the second set of equations are the normal equations for the least-squares problem (4.3) from the statement of the theorem. One sees this by noting that

$$\left(\underline{\mathbf{H}}_j^T\underline{\mathbf{H}}_j - \underline{\mathbf{H}}_j^T\mathbf{D}_j\mathbf{D}_j^T\underline{\mathbf{H}}_j\right) = \underline{\mathbf{H}}_j^T\left(\mathbf{I} - \mathbf{D}_j\mathbf{D}_j^T\right)\underline{\mathbf{H}}_j = \underline{\mathbf{H}}_j^T\mathbf{V}_{j+1}^T\left(\mathbf{I} - \mathbf{C}\mathbf{C}^T\right)\mathbf{V}_{j+1}\underline{\mathbf{H}}_j,$$

and that $\mathbf{C}\mathbf{C}^T = \mathbf{\Phi}$. For the right-hand side, one observes from the definition of $\mathbf{D}_j$ that

$$\underline{\mathbf{H}}_j^T\mathbf{D}_j\left(\mathbf{D}_j^T\mathbf{V}_{j+1}^T + \mathbf{F}_j^T\mathbf{C}_j^T\right) = \mathbf{V}_j^T\mathbf{A}^T\mathbf{C}\mathbf{C}^T\left(\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T + \mathbf{C}_j\mathbf{C}_j^T\right).$$

As we have seen in (4.4),

$$\mathbf{C}^T\left(\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T + \mathbf{C}_j\mathbf{C}_j^T\right) = \left[\left(\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T + \mathbf{C}_j\mathbf{C}_j^T\right)\mathbf{C}\right]^T = \mathbf{C}^T,$$

which means the right-hand side of the second equation of (4.5) can be simplified as $\underline{\mathbf{H}}_j^T\mathbf{V}_{j+1}^T\left(\mathbf{I} - \mathbf{\Phi}\right)\mathbf{r}_0$. Thus, we can rewrite (4.5) yielding

$$\begin{aligned}\mathbf{z}_j &= \mathbf{C}^T\mathbf{r}_0 - \mathbf{D}_j^T\underline{\mathbf{H}}_j\mathbf{y}_j \quad \text{and}\\ \underline{\mathbf{H}}_j^T\mathbf{V}_{j+1}^T\left(\mathbf{I} - \mathbf{\Phi}\right)\mathbf{V}_{j+1}\underline{\mathbf{H}}_j\mathbf{y}_j &= \underline{\mathbf{H}}_j^T\mathbf{V}_{j+1}^T\left(\mathbf{I} - \mathbf{\Phi}\right)\mathbf{r}_0.\end{aligned}$$

Observing the idempotency of projectors $(\mathbf{I}-\mathbf{\Phi}) = (\mathbf{I}-\mathbf{\Phi})^2$ completes the proof since $\mathbf{\Phi}$ being an orthogonal projector means it is symmetric. ∎

REMARK 4.2. We note that this result indicates that the matrices $\mathbf{C}_j$ and $\mathbf{F}_j$ are not needed to implement R^{3}GMRES, greatly simplifying the method, as it is no longer required to progressively orthogonalize $\mathbf{C}$ with respect to the Arnoldi vectors.

The final step in developing an efficient implementation of R^{3}GMRES is to rewrite and simplify equation (4.2) using the standard Givens-rotation-based progressive QR-factorization of $\underline{\mathbf{H}}_j$ which then enables the estimation of the residual norm without needing to compute the solution to equation (4.2) at each iteration. Unlike GMRES or the GCRO-variants of augmented methods, an exact residual norm is not available without computing the residual itself, which we would like to avoid.

THEOREM 4.3. *Let $\underline{\mathbf{H}}_j = \mathbf{Q}_j\underline{\mathbf{R}}_j$ be the QR-factorization obtained progressively using Givens rotations. Then we can represent the coefficient vectors $\mathbf{y}_j$ as the solution of the linear system*

$$\left(\mathbf{I}-\mathbf{M}_j\mathbf{M}_j^T\right)\mathbf{R}_j\mathbf{y}_j = \left\{\mathbf{Q}_j^T\left(\mathbf{V}_{j+1}^T\left(\mathbf{I}-\mathbf{\Phi}\right)\mathbf{r}_0\right)\right\}_{1:j},$$

where $\mathbf{M}_j = \left\{\mathbf{Q}_j^T\mathbf{D}_j\right\}_{1:j}\in\mathbb{R}^{j\times k}$. Furthermore, the residual norm satisfies

$$\begin{aligned}\left\|\left(\mathbf{I}-\mathbf{C}\mathbf{C}^T\right)\left(\mathbf{b}-\mathbf{A}\left(\mathbf{x}_0+\mathbf{V}_j\mathbf{y}_j\right)\right)\right\|^2 \\ \leq \left|\mathbf{e}_{j+1}^T\mathbf{Q}_j^T\mathbf{V}_{j+1}^T\mathbf{r}_0\right|^2 + \left\|\left(\mathbf{I}-\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T\right)\mathbf{r}_0\right\|^2,\end{aligned} \tag{4.6}$$

a bound which can be updated progressively.

Proof. Consider the QR-factorization $\underline{\mathbf{H}}_j = \mathbf{Q}_j\underline{\mathbf{R}}_j$, obtained progressively using Givens rotations. As has been observed in the derivation of GMRES in [27], we can write

$$\underline{\mathbf{H}}_j^T\underline{\mathbf{H}}_j = \underline{\mathbf{R}}_j^T\underline{\mathbf{R}}_j = \mathbf{R}_j^T\mathbf{R}_j.$$

With this, we can rewrite equation (4.2) as

$$\left(\mathbf{R}_j^T\mathbf{R}_j - \underline{\mathbf{R}}_j^T\mathbf{Q}_j^T\mathbf{D}_j\mathbf{D}_j^T\mathbf{Q}_j\underline{\mathbf{R}}_j\right)\mathbf{y}_j = \underline{\mathbf{R}}_j^T\mathbf{Q}_j^T\left(\mathbf{V}_{j+1}^T\left(\mathbf{I}-\mathbf{\Phi}\right)\mathbf{r}_0\right). \tag{4.7}$$

We assume that the Arnoldi process has not broken down and thus $\mathbf{R}_j$ is nonsingular. Thus, we can multiply the (4.7) by $\mathbf{R}_j^{-T}$, yielding

$$\left(\mathbf{R}_j - \underline{\mathbf{I}}_j^T\mathbf{Q}_j^T\mathbf{D}_j\mathbf{D}_j^T\mathbf{Q}_j\underline{\mathbf{R}}_j\right)\mathbf{y}_j = \underline{\mathbf{I}}_j^T\left(\mathbf{Q}_j^T\left(\mathbf{V}_{j+1}^T\left(\mathbf{I}-\mathbf{\Phi}\right)\mathbf{r}_0\right)\right). \tag{4.8}$$

Let $\mathbf{M}_j = \underline{\mathbf{I}}_j^T\mathbf{Q}_j^T\mathbf{D}_j\in\mathbb{R}^{j\times k}$. We can simplify (4.8) by substituting in $\mathbf{M}_j$, which yields

$$\left(\mathbf{I}-\mathbf{M}_j\mathbf{M}_j^T\right)\mathbf{R}_j\mathbf{y}_j = \left\{\mathbf{Q}_j^T\left(\mathbf{V}_{j+1}^T\left(\mathbf{I}-\mathbf{\Phi}\right)\mathbf{r}_0\right)\right\}_{1:j}. \tag{4.9}$$

If the rank-k outer product $\mathbf{M}_j\mathbf{M}_j^T$ does not have any unit eigenvalues, then $\left(\mathbf{I}-\mathbf{M}_j\mathbf{M}_j^T\right)$ is invertible. We note that this is indeed the case since (4.9) is derived from normal equations that have a unique solution in this case.

Recall from the proof of Theorem 4.1 that the solution to this linear system $\mathbf{y}_j$ is the minimizer of $\left\|\left(\mathbf{I}-\mathbf{C}\mathbf{C}^T\right)\left(\mathbf{V}_{j+1}\underline{\mathbf{H}}_j\mathbf{y}-\mathbf{r}_0\right)\right\|_2$. As $\mathbf{I}-\mathbf{C}\mathbf{C}^T$ is an orthogonal projection, its action either has no effect on the vector norm or it reduces the length. Thus we can estimate from above by disregarding the projector. Furthermore, this analysis should include the case that the Krylov subspace is range-restricted; thus $\mathbf{r}_0$ may not be in $\mathcal{R}\left(\mathbf{V}_{j+1}\right)$. As it has been

ETNA
Kent State University and
Johann Radon Institute (RICAM)

pointed out in (e.g., [20]), it suffices in this case to split the residual into $\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T\mathbf{r}_0$ and the part in the orthogonal complement and to consider the minimization only on the part in the Krylov subspace. Thus we can write

$$\begin{aligned}
&\left\|\left(\mathbf{I}-\mathbf{C}\mathbf{C}^T\right)\left(\mathbf{V}_{j+1}\underline{\mathbf{H}}_j\mathbf{y}-\mathbf{r}_0\right)\right\|^2 \\
&\quad\leq \left\|\left(\mathbf{V}_{j+1}\underline{\mathbf{H}}_j\mathbf{y}-\mathbf{r}_0\right)\right\|^2 = \left\|\left(\mathbf{V}_{j+1}\underline{\mathbf{H}}_j\mathbf{y}-\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T\mathbf{r}_0\right)-\left(\mathbf{I}-\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T\right)\mathbf{r}_0\right\|^2 \\
&\quad= \left\|\left(\mathbf{V}_{j+1}\underline{\mathbf{H}}_j\mathbf{y}-\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T\mathbf{r}_0\right)\right\|^2+\left\|\left(\mathbf{I}-\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T\right)\mathbf{r}_0\right\|^2 \\
&\quad= \left\|\left(\mathbf{Q}_j\underline{\mathbf{R}}_j\mathbf{y}-\mathbf{V}_{j+1}^T\mathbf{r}_0\right)\right\|^2+\left\|\left(\mathbf{I}-\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T\right)\mathbf{r}_0\right\|^2 \\
&\quad= \left\|\left(\underline{\mathbf{R}}_j\mathbf{y}-\mathbf{Q}_j^T\mathbf{V}_{j+1}^T\mathbf{r}_0\right)\right\|^2+\left\|\left(\mathbf{I}-\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T\right)\mathbf{r}_0\right\|^2.
\end{aligned}$$

The result follows from the same logic used to derive the GMRES residual monitoring strategy shown in, e.g., [27]. □

We note that if $\mathbf{w}_0=\mathbf{r}_0$, then

$$\left\|\left(\mathbf{I}-\mathbf{V}_{j+1}\mathbf{V}_{j+1}^T\right)\mathbf{r}_0\right\|=0.$$

If $\mathbf{w}_0=\mathbf{A}\mathbf{r}_0$, then one can progressively update and monitor this quantity by projecting $\mathbf{r}_0$ orthogonally away from $\mathcal{K}_j(\mathbf{A},\mathbf{A}\mathbf{r}_0)$.

We observe that the estimate of the residual norm is simply the residual norm one would obtain from applying non-augmented (range-restricted) GMRES to the problem. Thus, depending on the effectiveness of the augmentation, it will likely overestimate the true residual norm. However, the residual norm estimate (4.6) can be used in early iterations to avoid computing the solution and the residual until the estimate indicates convergence may be imminent. The strategy we advocate here is to use the ratio $\left\|\left(\mathbf{I}-\mathbf{C}\mathbf{C}^T\right)\mathbf{r}_0\right\|/\left\|\mathbf{r}_0\right\|$ as a scaling factor between the estimate of the norm and the actual norm. This scaling factor can be updated any time the code does an explicit residual computation, in the case we find that the estimate has falsely predicted convergence.

The matrix $\mathbf{M}_j$ can be constructed progressively using Givens rotations. We initialize $\mathbf{m}_1=\mathbf{d}_1$, reminding the reader that we are indexing the rows of $\mathbf{M}_j$. At iteration j, we set $\mathbf{m}_{j+1}=\mathbf{d}_{j+1}$ and use the jth Givens rotations to make the update

$$\begin{bmatrix}\mathbf{m}_j\\ \mathbf{m}_{j+1}\end{bmatrix}\leftarrow\begin{bmatrix}c_j & s_j\\ -s_j & c_j\end{bmatrix}\begin{bmatrix}\mathbf{m}_j\\ \mathbf{m}_{j+1}\end{bmatrix}.$$

We bring all this together to present a simplified implementation of R^{3}GMRES in Algorithm 2. Note that following from the strategy advocated by de Sturler [6], we compute the QR-factorization $\mathbf{A}\widehat{\mathbf{U}}=\mathbf{C}\mathbf{F}$, but we do not update $\mathbf{U}=\widehat{\mathbf{U}}\mathbf{F}^{-1}$. For $\mathbf{z}\in\mathbb{R}^k$, it is generally cheaper when expanding $\mathbf{U}\mathbf{z}$ to calculate $\widehat{\mathbf{U}}(\mathbf{F}^{-1}\mathbf{z})$. This is what we do in our implementation.

4.1. Comparison of implementations. We compare Algorithm 2 to [7, Algorithm 2] by studying their modifications to the common GMRES implementation upon which they are built, i.e., a Givens-rotation-based implementation as described in [27]. As in [7], we consider operations occurring inside the outermost loop. Inside of the main loop, both algorithms perform one matrix-vector product and an Arnoldi orthogonalization of each new basis vector. At the beginning of the algorithm, they perform many of the same or comparable initialization steps. According to the authors, Algorithm 2 in [7] performs $2k^3$ operations for additional Givens rotations per iteration since that method treats the minimization (4.1) directly. Additionally, obtaining an update of $\mathbf{C}_j$ at each iteration costs $2kn$ operations, and obtaining $\mathbf{F}_j$ costs $2k^2n$ at each iteration. Additionally, there are some lower-order costs. Thus,

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Algorithm 2: A simplified R^3GMRES implementation (with range restriction).

Input : $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{x}_0, \mathbf{b} \in \mathbb{R}^n$, $\mathbf{U} \in \mathbb{R}^{n\times k}$, $\varepsilon_{tol} > 0$

$[\mathbf{C}, \mathbf{F}] = \mathtt{QR}(\mathbf{AU})$
$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$; $\mathbf{w}_0 = \mathbf{A}\mathbf{r}_0$
$\gamma \leftarrow \left\|\mathbf{r}_0 - \mathbf{C}\mathbf{C}^T\mathbf{r}_0\right\| / \|\mathbf{r}\|_0$
$\mathbf{v}_1 \leftarrow \mathbf{w}_0 / \|\mathbf{w}_0\|_2$
$\mathbf{d}_1 \leftarrow \mathbf{v}_1^T\mathbf{C}$
$\mathbf{m}_1 \leftarrow \mathbf{d}_1$
$\mathbf{s}_1 = \mathbf{U}\left(\mathbf{F}^{-1}\left(\mathbf{C}^T\mathbf{r}_0\right)\right)$
for $i = 1, 2, \ldots, j$ **do**
    $\mathbf{v}_{i+1} \leftarrow \mathbf{A}\mathbf{v}_i$
    for $m = 1, 2, \ldots, i$ **do**
        $h_{mi} = \mathbf{v}_m^T\mathbf{v}_{i+1}$
        $\mathbf{v}_{i+1} \leftarrow \mathbf{v}_{i+1} - h_{mi}\mathbf{v}_m$
    end
    $h_{i+1,i} = \|\mathbf{v}_{i+1}\|_2$
    $\mathbf{v}_{i+1} = \mathbf{v}_{i+1}/h_{i+1,i}$
    $\mathbf{d}_{i+1} = \mathbf{v}_{i+1}^T\mathbf{C}$
    $\mathbf{m}_{i+1} \leftarrow \mathbf{d}_{i+1}$
    Apply previous rotations to jth column of $\underline{\mathbf{H}}_j$
    Obtain Givens sine and cosine s_j and c_j and updated $\mathbf{R}_j$
    Apply new rotations to update $\widehat{\mathbf{b}}_j = \mathbf{Q}_j^T\mathbf{V}_{j+1}^T\mathbf{r}_0$
    $\begin{bmatrix}\mathbf{m}_j \\ \mathbf{m}_{j+1}\end{bmatrix} \leftarrow \begin{bmatrix} c_j & s_j \\ -s_j & c_j\end{bmatrix}\begin{bmatrix}\mathbf{m}_j \\ \mathbf{m}_{j+1}\end{bmatrix}$
    if $\gamma \cdot |\widehat{\mathbf{b}}_j(j+1)| < \|\mathbf{r}_0\|\,\varepsilon_{tol}$ **then**
        Solve $\left(\mathbf{I} - \mathbf{M}_j\mathbf{M}_j^T\right)\mathbf{R}_j\mathbf{y}_j = \widehat{\mathbf{b}}_j(1:j)$
        Set $\mathbf{t} \leftarrow \mathbf{V}_j\mathbf{y}_j$
        Set $\mathbf{s}_2 = -\mathbf{U}\left(\mathbf{F}^{-1}\left(\mathbf{D}_j^T\mathbf{H}_j\mathbf{y}_j\right)\right)$
        Set $\mathbf{x} \leftarrow \mathbf{x}_0 + \mathbf{s}_1 + \mathbf{s}_2 + \mathbf{t}$; $\mathbf{r} \leftarrow \mathbf{b} - \mathbf{A}\mathbf{x}$
        if $\|\mathbf{r}\| < \|\mathbf{r}_0\|\,\varepsilon_{tol}$ **then**
            Exit loop and `return`
        else
            $\gamma \leftarrow \left\|\mathbf{r} - \mathbf{C}\mathbf{C}^T\mathbf{r}\right\| / \|\mathbf{r}\|$
    end
end

Algorithm 2 in [7] has a per-iteration cost above that of GMRES of roughly $2k(k^2 + kn + n)$ operations.

The formulation of Algorithm 2 allows us to discard many of these per-iteration operations. A comparable operation which is not discarded is the progressive building of $\mathbf{D}_j$, which costs nk operations. The update of $\mathbf{M}_j$ costs $2k$ operations. Thus, Algorithm 2 has a per-iteration cost above that of GMRES of roughly $k(n+2)$. Hence, the per-iteration cost of both algorithms above that of GMRES is $\mathcal{O}(n)$. The main difference is that the per iteration cost of Algorithm 2 above GMRES is linear in k whereas it is cubic for Algorithm 2 of [7]. Thus, we conclude that Algorithm 2 can accommodate a larger augmentation subspace with only linear growth in cost of additional operations.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

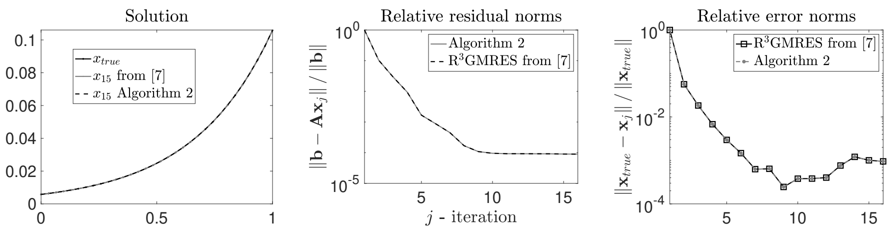

FIG. 5.1. *Experiment:* `deriv2()` *test.*

5. Numerical results. In this section, we demonstrate that Algorithm 2 produces approximations of the same quality as those produced by the version of the algorithm presented in [7] using code from the authors. The point here is not to compare the superiority of one version or the other, as neither code is optimized. Rather, as this note is laying out an alternative approach to the augmentation of unprojected Krylov subspaces, we demonstrate that our code delivers the same performance, verifying the alternative mathematical derivation in previous sections. We reproduce two experiments from [7] using `Regularization Tools` [16] with problem size $n = 256$. The noise vectors are generated from the normal distribution using `randn()`. For the experiments, we report the level of the noise relative to the size of the right-hand side, i.e., a relative noise level of 10^{-3} means that the 2-norm of the vector perturbing the right-hand side $\mathbf{b}_{true}$ is $10^{-3}\,\|\mathbf{b}_{true}\|$. All experiments are performed in Matlab R2020a, and we have established a repository [28] in which our code for Algorithm 2 is contained.

5.1. Experiment: `deriv2()` **test.** This reproduces the experiment in [7, Section 4.2], wherein augmentation is used to help encode known boundary conditions approximately so that the iteration focuses mostly on reconstructing the solution in the interior of the domain. The matrix is generated by the `deriv2()` function which produces a discretization of the Fredholm integral operator whose kernel is the Green's function of the second derivative operator. The relative noise level is 10^{-5}. Following [7, Section 4.2], we set $\mathcal{U} = \operatorname{span}\left\{\begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}^{\mathrm{T}}, \begin{bmatrix}1 & 2 & \cdots & n\end{bmatrix}^{\mathrm{T}}\right\}$. Results shown in Figure 5.1 demonstrate that the performance of the two implementations is virtually indistinguishable.

5.2. Experiment: `gravity()` **test—correctly localized discontinuity.** We generate the matrix $\mathbf{A}$ for this example using the `gravity()` function, which generates a discretization of a Fredholm integral operator of the first kind modeling a one-dimensional gravity surveying problem application posed on the interval $[0, 1]$. The relative noise level is 10^{-4}. We take the true solution produced by the function and introduce a discontinuity at $t = \frac{1}{2}$, as in [7, Section 4.3]. For this experiment, we assume we know the location of the discontinuity and set $\mathcal{U} = \operatorname{span}\left\{\begin{bmatrix}0 & 0 & \cdots & 0 & 1 & \cdots & 1\end{bmatrix}^{\mathrm{T}}\right\}$ to correctly encode this discontinuity. In Figure 5.2, we see that the two implementations perform identically.

5.3. Experiment: `gravity()` **test—incorrectly localized discontinuity.** We construct the same problem as in the previous experiment, but we move the discontinuity to a $t > \frac{1}{2}$. However, we encode the discontinuity incorrectly using the same $\mathcal{U}$ as in the previous experiment. In Figure 5.3, we observe that both implementations again perform identically. Furthermore, one sees that the minimization process reduces the influence of the falsely-placed discontinuity encoded by $\mathcal{U}$ while trying to fit the true discontinuity. This has been noted in [7, Section 4.3] as a possible advantage in augmenting an unprojected Krylov subspace for solving an ill-posed problem, as the incorrectly-chosen $\mathcal{U}$ does not influence

ETNA
Kent State University and
Johann Radon Institute (RICAM)

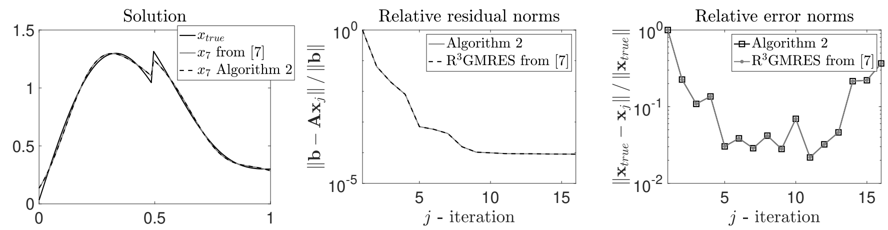

FIG. 5.2. *Experiment:* `gravity()` *test—correctly localized discontinuity.*

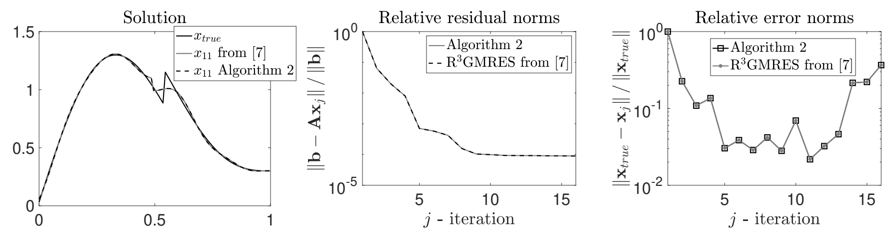

FIG. 5.3. *Experiment:* `gravity()` *test—incorrectly localized discontinuity.*

which Krylov subspace is built. We contrast this with the best reconstruction produced by an augmented iterative solver using a projected Krylov subspace, $\mathcal{K}\left(\left(\mathbf{I}-\boldsymbol{\Phi}\right)\mathbf{A}, \left(\mathbf{I}-\boldsymbol{\Phi}\right)\mathbf{r}_0\right)$, using a GCRO-type code. In Figure 5.4, we see that the method at its best still emphasizes the incorrectly localized discontinuity.

6. Discussion. The main goal in this note is to demonstrate that augmented unprojected Krylov subspace methods fit into the same framework from [29] enabling a simpler implementation in the style of a GCRO-DR-type method. This leads us to observe that the R^{3}GMRES method is closely related to the augmentation strategy from [26]. With that perspective, we show one can actually approximate the solution to a projected subproblem and project the approximation to obtain the part from the augmented subspace. The benefit when applying this to the R^{3}GMRES method is that we no longer need to progressively maintain an orthonormal basis for the full sum subspace $\mathbf{A}\left(\mathcal{U}+\mathcal{V}_j\right)$.

The numerical experiments we showed follow from what was done in [7], focusing on instances wherein one wants to enforce that the solution has an a priori known structure but accommodate the possibility that this knowledge is flawed. We contrasted this with the performance of a GCRO-type method to show how for an ill-posed problem, an augmented method with a projected Krylov subspace can over-emphasize the bad knowledge to an extent that it cannot recover due to the ill-posedness of the problem.

However, it should be noted that GCRO-based augmentation/recycling methods still exhibit superior performance when it comes to the acceleration of convergence for complicated, large-scale problems. Rather, this work highlights that it can be important to distinguish between "trustworthy" and "untrustworthy" information when using augmentation methods, particularly for ill-posed problems. A future path to explore would be to consider mixing the two strategies more generally for situations when one has both trustworthy and untrustworthy/corrupted information that one wishes to use without it corrupting the behavior of the solver.

Acknowledgments. The author thanks Per Christian Hansen for an interesting discussion about this topic back when we could go to conferences in person and for sending the author an

ETNA
Kent State University and
Johann Radon Institute (RICAM)

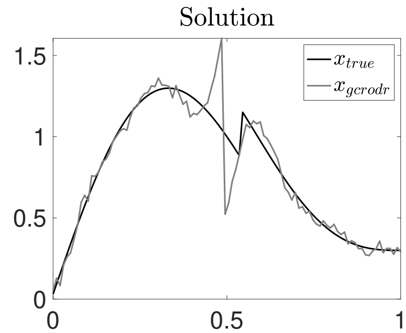

FIG. 5.4. `gravity()` *test—incorrectly localized discontinuity with projected Krylov subspace.*

implementation of R^{3}GMRES to validate against. The author also thanks the two anonymous referees for their helpful comments and suggested edits to tighten up the exposition of the manuscript.

REFERENCES

[1] J. BAGLAMA AND L. REICHEL, *Augmented GMRES-type methods*, Numer. Linear Algebra Appl., 14 (2007), pp. 337–350.

[2] ———, *Decomposition methods for large linear discrete ill-posed problems*, J. Comput. Appl. Math., 198 (2007), pp. 332–343.

[3] J. CHUNG AND S. GAZZOLA, *Computational methods for large-scale inverse problems: a survey on hybrid projection methods*, Preprint on arXiv, 2021. https://arxiv.org/abs/2105.07221

[4] E. DE STURLER, *Nested Krylov methods based on GCR*, J. Comput. Appl. Math., 67 (1996), pp. 15–41.

[5] ———, *Truncation strategies for optimal Krylov subspace methods*, SIAM J. Numer. Anal., 36 (1999), pp. 864–889.

[6] ———, Private communication, 2020.

[7] Y. DONG, H. GARDE, AND P. C. HANSEN, *R^{3}GMRES: including prior information in GMRES-type methods for discrete inverse problems*, Electron. Trans. Numer. Anal., 42 (2014), pp. 136–146. http://etna.ricam.oeaw.ac.at/vol.42.2014/pp136-146.dir/pp136-146.pdf

[8] L. DYKES, F. MARCELLÁN, AND L. REICHEL, *The structure of iterative methods for symmetric linear discrete ill-posed problems*, BIT, 54 (2014), pp. 129–145.

[9] J. ERHEL AND F. GUYOMARC'H, *An augmented conjugate gradient method for solving consecutive symmetric positive definite linear systems*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1279–1299.

[10] A. GAUL, *Recycling Krylov Subspace Methods for Sequences of Linear Systems: Analysis and Applications*, PhD. Thesis, Fakultät II-Mathematik und Naturwissenschaften, Technischen Universität Berlin, Berlin, 2014.

[11] A. GAUL, M. H. GUTKNECHT, J. LIESEN, AND R. NABBEN, *A framework for deflated and augmented Krylov subspace methods*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 495–518.

[12] S. GAZZOLA AND M. S. LANDMAN, *Krylov methods for inverse problems: surveying classical, and introducing new, algorithmic approaches*, GAMM-Mitt., 43 (2020), Art. e202000017, 31 pages.

[13] M. H. GUTKNECHT, *Spectral deflation in Krylov solvers: a theory of coordinate space based methods*, Electron. Trans. Numer. Anal., 39 (2012), pp. 156–185. http://etna.ricam.oeaw.ac.at/vol.39.2012/pp156-185.dir/pp156-185.pdf

[14] ———, *Deflated and augmented Krylov subspace methods: a framework for deflated BiCG and related solvers*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 1444–1466.

[15] M. HANKE, *Conjugate Gradient Type Methods for Ill-Posed Problems*, Longman Scientific, Harlow, 1995.

[16] P. C. HANSEN, *Regularization tools version 4.0 for Matlab 7.3*, Numer. Algorithms, 46 (2007), pp. 189–194.

[17] J. JIANG, J. CHUNG, AND E. DE STURLER, *Hybrid projection methods with recycling for inverse problems*, SIAM J. Sci. Comput. 43 (2021), pp. S146–S172.

[18] R. B. MORGAN, *GMRES with deflated restarting*, SIAM J. Sci. Comput., 24 (2002), pp. 20–37.

[19] A. NEUMAN, L. REICHEL, AND H. SADOK, *Algorithms for range restricted iterative methods for linear discrete ill-posed problems*, Numer. Algorithms, 59 (2012), pp. 325–331.

[20] ———, *Implementations of range restricted iterative methods for linear discrete ill-posed problems*, Linear Algebra Appl., 436 (2012), pp. 3974–3990.

[21] M. L. PARKS, E. DE STURLER, G. MACKEY, D. D. JOHNSON, AND S. MAITI, *Recycling Krylov subspaces for sequences of linear systems*, SIAM J. Sci. Comput., 28 (2006), pp. 1651–1674.

[22] M. L. PARKS, K. M. SOODHALTER, AND D. B. SZYLD, *A block recycled GMRES method with investigations into aspects of solver performance*, Preprint on arxiv, 2016. https://arxiv.org/abs/1604.01713

[23] R. RAMLAU, K. M. SOODHALTER, AND V. HUTTERER, *Subspace recycling-based regularization methods*, SIAM J. Matrix Anal. Appl., 42, (2021), pp. 1480–1505.

[24] R. RAMLAU AND B. STADLER, *An augmented wavelet reconstructor for atmospheric tomography*, Electron. Trans. Numer. Anal., 54 (2021), pp. 256–275. http://etna.ricam.oeaw.ac.at/vol.54.2021/pp256-275.dir/pp256-275.pdf

[25] L. REICHEL AND Q. YE, *Breakdown-free GMRES for singular systems*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 1001–1021.

[26] Y. SAAD, *Analysis of augmented Krylov subspace methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 435–449.

[27] Y. SAAD AND M. H. SCHULTZ, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[28] K. M. SOODHALTER, *kirkmsoodhalter/r3gmres-simplified: R3gmres simplified implementation*, 2021, source at https://doi.org/10.5281/ZENODO.4975990.

[29] K. M. SOODHALTER, E. D. STURLER, AND M. E. KILMER, *A survey of subspace recycling iterative methods*, GAMM-Mitt., 43 (2020), Art. e202000016, 29 pages.