# Sketched GCRODR and its Convergence Analysis

*Eric de Sturler and Fei Xue*

Abstract

We develop a sketched version of the GCRODR algorithm for the solution of a sequence of linear systems. The recycling approach in GCRODR with an approximate invariant subspace allows us to derive upperbounds on the convergence of GCRODR based on the field of values of the projected system (see below). We extend this convergence result to upperbounds on the convergence of a sketched GCRODR (S-GCRODR). The bounds for S-GCRODR deteriorate from those for GCRODR as a function of the subspace embedding distortion $\epsilon$, and we provide expressions for this relation.

Sketching offers the opportunity to substantially reduce the high orthogonalization cost in long-recurrence solvers like GMRES. Several approaches have been explored. Balabanov and Grigori [1] replace the inner products in the orthogonalization by sketched inner products, replacing an orthogonal projection by a oblique projection (but typically close to orthogonal), which maintains the stability of the Arnoldi process. While they demonstrate good performance improvements on HPC architectures, the approach does not reduce the computational complexity, $O(nm^2)$ for $m$ iterations with $A \in \mathbb{C}^{n\times n}$. On the other hand, Nakatsukasa and Tropp [5] generate the Krylov space with truncated Arnoldi and use sketching for the LS solution of the resulting, potentially very ill-conditioned, system. This approach has the significant advantage that it drastically reduces the computational complexity to $O(nm \log m)$ for $m$ iterations. However, the severe ill-conditioning of the basis vectors typically leads to some deterioration of the convergence. We propose an efficient and convergence-wise effective combination of the two approaches.

We consider the solution of a sequence of linear systems, $A^{(j)}x^{(j)} = b^{(j)}$, where $A \in \mathbb{C}^{n\times n}$, and where the matrices change slowly. We aim for robustness and reduced iterations as well as a significant reduction in the average cost per iteration. Recycling Krylov subspaces from previous linear solves can drastically reduce the total number of iterations, which suggests that the approximate orthogonalization by sketching of new Krylov vectors against the recycle space is important. This introduces only a linear cost in the number of iterations. In addition, we substantially reduce the computational complexity by using only selective orthogonalization with a fixed number of orthogonalizations when we extend the (augmented) Krylov search space and solve the least squares problem in a sketched fashion following the approach proposed in [5].

**GCRODR and S-GCRODR** Consider a recycle space of dimension $k$, defined by (range) $R(U)$, where $U \in \mathbb{C}^{n\times k}$ such that (for convenience) $C = AU$ has orthonormal columns, $C^*C = I$. We define the (orthogonal) projection $\Phi = CC^*$. We also define $C_\perp$ such that the matrix $[C\ C_\perp] \in \mathbb{C}^{n\times n}$ is unitary. We assume here, for simplicity, that $U$ has been selected such that $R(U)$ is a low accuracy approximation (see below) to an invariant subspace with eigenvalues near the origin. As shown in [7], using the recycle space $R(U)$, we can update the initial solution, $\widetilde{x}_0$, and residual, $\widetilde{r}_0$, as $x_0 = \widetilde{x}_0 + UC^*\widetilde{r}_0$, and $r_0 = (I-\Phi)\widetilde{r}_0$, and subsequently solve the *projected system* $(I-\Phi)Az = (I-\Phi)\widetilde{r}_0$ with GMRES. For this (consistent) system, the right hand side $(I-\Phi)\widetilde{r}_0 \in R(C)^\perp = R(C_\perp)$ and $(I-\Phi)Az : R(C_\perp) \to R(C_\perp)$. So, we can analyze the convergence for GMRES for the linear operator $(I-\Phi)A$ over the space $R(C_\perp)$.

After defining a sketching matrix $S \in \mathbb{C}^{s\times n}$, which provides an $\ell_2$ embedding of a suitable vector space $\mathcal{V}$, which contains the right hand side or residual, the $R(C)$, and a suitable Krylov space, we

let $SC = YR_Y$ and $S^*Y = QR_Q$ be reduced QR decompositions. In S-GCRODR the orthogonal projection $I-\Phi$ in GCRODR is replaced by the (oblique) projection $I-\widehat{\Phi}$, with range $R(\widehat{\Phi}) = R(C)$ and null space $N(\widehat{\Phi}) = R(Q)^\perp = R(Q_\perp)$, where $[Q\ Q_\perp]$ is a unitary matrix. This implies that $(I-\widehat{\Phi}) = Q_\perp(C_\perp^* Q_\perp)^{-1}C_\perp^*$. After computing the updates to the initial guess and residual, S-GCRODR solves the *projected system* $(I-\widehat{\Phi})Az = (I-\widehat{\Phi})\widetilde{r}_0$ using GMRES.

**Convergence** We give bounds on the convergence for GCRODR while recycling an approximate invariant subspace and compare these with convergence bounds for the sketched version, S-GCRODR, recycling the same invariant subspace. We show that the convergence bounds for S-GCRODR can deteriorate due the oblique projection; however, the deterioration can be bounded in terms of the embedding subspace distortion $\epsilon$.

We can analyze the convergence of GCRODR by considering convergence bounds for GMRES for the linear operator $(I-\Phi)A$ restricted to the space $R(C_\perp)$, which can derived using the field of values (FOV) [4, 3] of $C_\perp^* A C_\perp$. Now let $A$ have the block Schur decomposition (with unitary $[V\ V_\perp] \in \mathbb{C}^{n\times n}$)

$$A \;=\; [V\ V_\perp] \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} [V\ V_\perp]^*, \tag{1}$$

where the eigenvalues of $T_{11}$ are near the origin (possibly surrounding the origin) and $\|T_{11}\|_2$ is small, and the eigenvalues of $T_{22}$ are further away in the right half plane, and let $\|(I-\Phi)V\|_2 = \delta < 1$. In the derivation of FOV bounds, we use the following notation. We use calligraphic script to denote sets: $\mathcal{F}(T_{11})$ denotes the field of values of $T_{11}$, $\mathcal{F}(T_{22})$ denotes the field of values of $T_{22}$ and $\mathcal{D}$ denotes the unit disk. Set addition is defined in the usual way, and for a scalar $\tau$ and set $\mathcal{S}$, the set $\tau\mathcal{S}$ is defined as $\tau\mathcal{S} = \{\tau x \mid x \in \mathcal{S}\}$ and $[\tau_1,\tau_2]\mathcal{S} = \{\tau x \mid x \in \mathcal{S} \text{ and } \tau \in [\tau_1,\tau_2]\}$. We can then bound the FOV of the linear operator $(I-\Phi)A$ restricted to the space $R(C_\perp)$, $\mathcal{F}(\mathcal{C}_\perp^* \mathcal{A}\mathcal{C}_\perp)$ as

$$\mathcal{F}(\mathcal{C}_\perp^* \mathcal{A}\mathcal{C}_\perp) \;\subset\; [1-\delta^2, 1]\mathcal{F}(T_{22}) + [0,\delta^2]\mathcal{F}(T_{11}) + \delta(1-\delta^2)^{1/2}\|T_{12}\|_2\mathcal{D}. \tag{2}$$

This equation shows that even for $\delta$ not very small, say $\delta = 10^{-2}$ (which can be achieved with modest effort [6]), $\mathcal{F}(\mathcal{C}_\perp^* \mathcal{A}\mathcal{C}_\perp)$ is only slightly large than $\mathcal{F}(T_{22})$, unless $\|T_{12}\|_2$ is (relatively) large. We can now bound the convergence of GCRODR for $A$ with the recycle space $R(U)$ using the FOV convergence bounds for GMRES with the FOV bounds from (2).

We can bound the convergence of S-GCRODR in a similar fashion as for GCRODR using bounds on the FOV of $(I-\widehat{\Phi})A$ restricted to the space $R(Q_\perp)$, that is the set $\{z^*(I-\widehat{\Phi})Az \;:\; z = Q_\perp\zeta, \zeta \in \mathbb{C}^{n-k}, \|\zeta\|_2 = 1\}$, which is also given by $\mathcal{F}(Q_\perp^* Q_\perp (C_\perp^* Q_\perp)^{-1} C_\perp^* A Q_\perp) = \mathcal{F}(\ (C_\perp^* Q_\perp)^{-1} C_\perp^* A Q_\perp\ )$.

To understand the relation between the FOV bounds for GCRODR and S-GCRODR, we consider the singular values of $C_\perp^* Q_\perp$. We assume $k \ll n$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ be the singular values of $C^*Q$. Then we can derive the singular values of $C_\perp^* Q_\perp$ from the CS-decomposition of $[C\ C_\perp]^*[Q\ Q_\perp]$: $\sigma(C_\perp^* Q_\perp) \in \{1, \lambda_1, \ldots, \lambda_k\}$. Furthermore, we can prove, based on the $\epsilon$-embedding, that $\lambda_k \geq \sqrt{(1-\epsilon)/(1+\epsilon)}$, and therefore, for $\epsilon \to 0$, $R(Q_\perp) \to R(C_\perp)$. This in turn implies that $(C_\perp^* Q_\perp)^{-1} C_\perp^* A Q_\perp \to C_\perp^* A C_\perp$, and hence the FOVs that govern the convergence bounds for GCRODR and S-GCRODR get closer and closer as $\epsilon$ becomes small.

We can describe the dependence of $\mathcal{F}(\ (C_\perp^* Q_\perp)^{-1} C_\perp^* A Q_\perp\ )$ on $\epsilon$ in substantial detail by deriving detailed expressions of the type (for unit vectors $\zeta$)

$$\zeta^* Q_\perp^* Q_\perp (C_\perp^* Q_\perp)^{-1} C_\perp^* A Q_\perp \zeta \;=\; \begin{pmatrix} \eta_1(\epsilon) \\ \eta_2(\epsilon) \end{pmatrix}^* \begin{pmatrix} T_{11} & T_{12} \\ O & T_{22} \end{pmatrix} \begin{pmatrix} \eta_1(\epsilon) \\ \eta_2(\epsilon) \end{pmatrix}, \tag{3}$$

and analyze how close bounds for (3) are to (2) as a function of $\epsilon$. These bounds clarify how the convergence of S-GCRODR may deteriorate as a function of the distortion parameter $\epsilon$ as a consequence of how far the oblique projection $I - \widehat{\Phi}$ deviates from the orthogonal projection $I - \Phi$.

**A Numerical Experiment** We present one set of numerical results to compare several sketched variants of GMRES and GCRODR. The results are derived from solving the following nonlinear Helmholtz equation on the 2D domain $\Omega = (0,1) \times (0,1)$,

$$\begin{cases} \Delta u + \kappa^2(1 + \epsilon|u|^2)u = 0, \\ u_x + i\kappa u = 2i\kappa, \quad \text{at } x = 0, y \in (0,1) \\ u_x - i\kappa u = -i\kappa, \quad \text{at } x = 1, y \in (0,1) \\ \text{periodic boundary condition at } y = 0, 1, x \in (0,1), \end{cases} \tag{4}$$

using Anderson acceleration (AA). We take $\epsilon = 0.40$ and $\kappa = 12$. We discretize $\Omega$ using a uniform mesh with $n+1$ equispaced nodes in the $x$ and in the $y$ directions, respectively. We also use the standard 2nd order finite difference to approximate the Laplacian operator, and use ghost nodes at the left ($x = 0$) and the right ($x = 1$) boundaries. We let $n = 512$ so that the number of elements in the $u$ vectors is $n(n+1) = 262656$. To set up the corresponding nonlinear system, define $I_x = I_{n+1}$, $I_y = I_n$, $D_{2x} = \text{tridiag}(\mathbf{1}, -\mathbf{2}, \mathbf{1}) \in \mathbb{R}^{(n+1)\times(n+1)}$, except that $D_{2x}(1,1) = D_{2x}(n+1, n+1) = 2(-1 + i\kappa h)$, and $D_{2x}(1,2) = D_{2x}(n+1, n) = 2$, $D_{2y} = \text{tridiag}(\mathbf{1}, -\mathbf{2}, \mathbf{1}) \in \mathbb{R}^{n\times n}$, except that $D_{2y}(1,n) = D_{2y}(n,1) = 1$, $F(u) = \frac{1}{h^2}(I_y \otimes D_{2x} + D_{2y} \otimes I_x) + \kappa^2 \text{diag}(1 + \epsilon|u|^2) - f_{bdy}$, where $f_{bdy} = \mathbf{1}_n \otimes [\frac{2(2)i\kappa}{h}; \mathbf{0_{n-1}}; \frac{2(-1)i\kappa}{h}]$, so that the nonlinear system is $F(u)u = 0$. To define the Picard iteration, we let the squared term of $u$ be the current iterate $u^{(k)}$ and solve for the next iterate $u^{(k+1)}$. That is, at each step we solve the linear system

$$\left(\frac{1}{h^2}(I_y \otimes D_{2x} + D_{2y} \otimes I_x) + \kappa^2 \text{diag}(1 + \epsilon|u^{(k)}|^2)\right) u^{(k+1)} = f_{bdy} \tag{5}$$

for $u^{(k+1)}$. The initial vector $u^{(0)}$ is the vectorization of $u(x,y) = e^{i(2\pi y + \kappa x)}$ on the mesh. To set up AA, we let the damping parameter be 1, the optimization involve all previous iterates $u^{(k)}$, and the iteration is terminated when $\|u^{(k+1)} - u^{(k)}\|_\infty \leq 10^{-6}$.

At each step of AA, the linear system (5) is solved by the following methods, and a new ILUTP preconditioner is constructed using approximate minimum degree ordering and drop tolerance 0.002. We compare GMRES(120), the sketched version S-GMRES(120) as proposed in [5], (standard) GCRODR(120,20) and the sketched version S-GCRODR(120,20) discussed above, and two versions of the method GMRES-SDR(120,20) proposed in the recent paper [2], where the authors combine a sketched version of GMRES with deflated restarting. This approach differs from S-GCRODR in that the authors apply the deflated restarting by augmenting the search space with the deflation vectors, using truncated/selective orthogonalization when generating new Krylov search directions, and then using sketching to solve the least squares problem over both deflation and new Krylov vectors. In this approach, the new Krylov space that extends the solution search space is not generated (approximately) orthogonal to (the image under $A$ of) the recycle space. This may lead to less effective search spaces and hence a reduced convergence rate. On the other hand, for the same total search space dimension in a cycle, it leads to a further reduction in complexity compared with the method we propose.

In Table 1, for each linear solver, we give the total and average runtime and number of preconditioned matrix-vector products for solving the sequence of linear systems (5) arising from Anderson

Table 1: Total and average numbers of preconditioned matrix-vector products and runtimes for several methods solving the sequence of linear systems (5) arising from Anderson acceleration for a nonlinear Helmholtz equation.

| | GMRES(120) | GCRO-DR(120,20) | S-GMRES(120) | S-GCRO-DR(120,20) | (m) GMRES-SDR(120,20) | (s) GMRES-SDR(120,20) |
|---|---|---|---|---|---|---|
| matvecs | 9014 (361) | 2794 (121) | 10319 (382) | 2819 (123) | 10144 (423) | 6037 (232) |
| time (secs) | 1374.3 (55.0) | 343.8 (14.9) | 691.3 (25.6) | 183.2 (8.0) | 622.5 (25.9) | 393.4 (15.1) |

acceleration for the nonlinear system (4). For all linear solvers, we let the maximum dimension of the subspace be $m = 120$ and let the recycle space dimension be $k = 20$. Due to the irregular convergence behavior of Anderson acceleration with the linear systems (5) solved *approximately*, it takes Anderson acceleration a slightly different number of steps to satisfy the stopping criterion $\|u^{(k+1)} - u^{(k)}\|_\infty \leq 10^{-6}$ for each solver. AA based on GCRO-DR and S-GCRO-DR takes 23 (the fewest) steps, whereas AA based on S-GMRES takes 27 (the most) steps to converge. In the column '(m) GMRES-SDR(120,20)' we report results when GMRES-SDR recycles search spaces from one linear system to the next, whereas under the column '(s) GMRES-SDR(120,20)' we report results with GMRES-SDR starting each linear system without a recycle space (which seems to work better). Finally, we note that, while for this system S-GCRODR is the clear winner, by a large margin, in terms of the runtime, for other test problems GMRES-SDR was competitive and sometimes faster.

# References


[1] O. Balabanov and L. Grigori, *Randomized Gram–Schmidt process with application to GMRES*, SIAM J. Sci. Comput., 44 (2022), pp. A1450–A1474.

[2] L. Burke, S. Güttel, and K. Soodhalter, *GMRES with randomized sketching and deflated restarting*, arXiv:2311.14206, arXiv, 2023.

[3] M. Embree, *Extending Elman's Bound for GMRES*, arXiv:2312.15022v1, arXiv, 2023.

[4] A. Greenbaum, *Iterative methods for solving linear systems*, SIAM 1997.

[5] Y. Nakatsukasa and J. A. Tropp, *Fast and accurate randomized algorithms for linear systems and eigen- value problems*, SIAM J. Matrix Anal. Appl., 45 (2024), pp. 1183–1214.

[6] 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.

[7] K.M. Soodhalter, E. de Sturler, and M. E. Kilmer, *A survey of subspace recycling iterative methods*, GAMM-Mitteilungen 43 (2020), p. e202000016.