Electronic Transactions on Numerical Analysis.
Volume 53, pp. 541–561, 2020.
Copyright © 2020, Kent State University.
ISSN 1068–9613.
DOI: 10.1553/etna_vol53s541

KRYLOV TYPE METHODS FOR LINEAR SYSTEMS EXPLOITING PROPERTIES OF THE QUADRATIC NUMERICAL RANGE[*]

ANDREAS FROMMER[†], BIRGIT JACOB[†], KARSTEN KAHL[†], CHRISTIAN WYSS[†], AND IAN ZWAAN[†]

Abstract. The quadratic numerical range $W^2(A)$ is a subset of the standard numerical range of a linear operator, which still contains its spectrum. It arises naturally in operators that have a 2×2 block structure, and it consists of at most two connected components, none of which necessarily convex. The quadratic numerical range can thus reveal spectral gaps, and it can in particular indicate that the spectrum of an operator is bounded away from 0.

We exploit this property in the finite-dimensional setting to derive Krylov subspace-type methods to solve the system $Ax = b$, in which the iterates arise as solutions of low-dimensional models of the operator whose quadratic numerical range is contained in $W^2(A)$. This implies that the iterates are always well-defined and that, as opposed to standard FOM, large variations in the approximation quality of consecutive iterates are avoided, although 0 lies within the convex hull of the spectrum. We also consider GMRES variants that are obtained in a similar spirit. We derive theoretical results on basic properties of these methods, review methods on how to compute the required bases in a stable manner, and present results of several numerical experiments illustrating improvements over standard FOM and GMRES.

Key words. quadratic numerical range, full orthogonalization method (FOM), generalized minimal residual method (GMRES), linear systems, projection methods, two-level orthogonal Arnoldi method

AMS subject classifications. 15A60, 47A12, 65F10, 65N55

1. Introduction.

It is well known that Krylov subspace methods for a linear system $Ax = b$ with a nonsingular matrix $A \in \mathbb{C}^{n\times n}$ might converge slowly, or even diverge, or fail in situations where 0 lies in the “interior” of the spectrum $\operatorname{spec}(A)$ of A. Specifically, if 0 is contained in the numerical range (or field of values) of A—a convex set containing $\operatorname{spec}(A)$—we know that methods based on a Galerkin variational characterization, like the full orthogonalization method FOM, can fail due to the non-existence of certain iterates. This manifests itself numerically by huge variations in magnitude and associated stability problems. In methods that are based on residual minimization, like the generalized minimal residual method GMRES, stagnation can occur in such cases. Such stagnation might occur during a large number of iterations, which is highly undesirable, but it can also be more benign. For example, the MINRES method, which is mathematically equivalent to GMRES for Hermitian matrices A, typically achieves good progress in every other iteration when A is indefinite and not too ill-conditioned. Related to this, classical convergence theory for Krylov subspace methods, in particular for the non-Hermitian case, typically assumes that 0 is not contained in the numerical range and then gets quantitative results on convergence speed in which the distance of the numerical range to 0 enters as a parameter; see, e.g., [3, 20, 21] and the discussion and references in the books [12, 19].

In this paper we study modifications of the FOM method, and also of GMRES, which converge stably and smoothly when the *quadratic numerical range*, a subset of the standard numerical range, splits into two parts that do not contain 0. The quadratic numerical range arises naturally for matrices that have a canonical 2×2 block structure. Analogously to standard Krylov subspace methods, these modifications are also based on projections. By projecting onto a larger space than the Krylov subspace, we manage to preserve the gap in the quadratic numerical range and thus shield the projected matrices away from singularity. At the same time, we do not require more matrix-vector multiplications than in standard Krylov

[*] Received December 24, 2019. Accepted August 16, 2020. Published online on November 27, 2020. Recommended by Jocelyne Erhel. This work was supported in part by Deutsche Forschungsgemeinschaft through the collaborative research centre SFB-TRR55.

[†] Department of Mathematics, Bergische Universität Wuppertal, Gaußstraße 20, 42097 Wuppertal, Germany ({frommer,jacob,kkahl,wyss,zwaan}@math.uni-wuppertal.de).

subspace methods, namely just one per iteration. While the projection technique underlying our approach has already been applied in other contexts, like, for example, in symplectic model order reduction techniques (see, e.g., [17]), we note that its use for iterative linear solvers appears to be new as is the study of its connection to the quadratic numerical range.

This paper is organized as follows: Section 2 reviews those properties of the numerical range and the FOM and GMRES method that are important for this paper. Section 3 first introduces the quadratic numerical range and then develops the new modified projection methods termed quadratic FOM and quadratic GMRES. This section also contains first elements of an analysis. In Section 4 we then discuss how the new methods can be realized as efficient algorithms before we give some numerical examples in Section 5.

2. Numerical range and FOM. Regardless of the dimension n, we will always denote by $\langle\cdot,\cdot\rangle$ the standard sesquilinear inner product on $\mathbb{C}^n$ and by $\|\cdot\|$ the associated norm. For a linear operator $A \in \mathbb{C}^{n\times n}$ the numerical range (or field of values) $W(A)$ is the set of all its Rayleigh quotients

$$W(A) = \left\{ \frac{\langle Ax, x\rangle}{\langle x, x\rangle} : x \in \mathbb{C}^n, x \neq 0 \right\} = \{\langle Ax, x\rangle : x \in \mathbb{C}^n, \|x\| = 1\}.$$

$W(A)$ is a compact convex set (see, e.g., [8]), which contains the spectrum $\operatorname{spec}(A)$. If A is normal, that is if $A^*A = AA^*$, then $W(A)$ is actually the convex hull of $\operatorname{spec}(A)$. For non-normal A, the numerical range $W(A)$ can be much larger than the convex hull of the spectrum. If for some $m \le n$ the matrix $V = [v_1 \mid \cdots \mid v_m] \in \mathbb{C}^{n\times m}$ is an orthonormal matrix, i.e., $V^*V = I_m$, the identity on $\mathbb{C}^m$, then the numerical range of the "projected" matrix $V^*AV \in \mathbb{C}^{m\times m}$ is contained in that of A, since for all $y \in \mathbb{C}^m, y \neq 0$ we have $\langle y, y\rangle = \langle Vy, Vy\rangle$ and thus

$$\frac{\langle V^*AVy, y\rangle}{\langle y, y\rangle} = \frac{\langle AVy, Vy\rangle}{\langle y, y\rangle} = \frac{\langle AVy, Vy\rangle}{\langle Vy, Vy\rangle} \in W(A).$$

For future use we state this observation as a lemma.

LEMMA 2.1. *Let $A \in \mathbb{C}^{n\times n}$ be arbitrary and let $V \in \mathbb{C}^{n\times m}$ be orthonormal. Then*

$$W(V^*AV) \subseteq W(A).$$

We continue by summarizing the basic properties of two Krylov subspace methods, namely FOM [18] and GMRES [20], that are relevant for this work. We refer to the textbooks [2, 5, 12, 19, 22] for a summary of the plethora of further properties of these two methods and of Krylov subspace methods in general.

A Krylov subspace method for solving the linear system

$$Ax = b, \quad A \in \mathbb{C}^{n\times n}, b \in \mathbb{C}^n,$$

takes its k-th iterate from the affine subspace $x^{(0)} + \mathcal{K}^{(k)}(A, r^{(0)})$, where $r^{(0)} = b - Ax^{(0)}$ and

$$\mathcal{K}^{(k)}(A, r^{(0)}) = \operatorname{span}\{r^{(0)}, Ar^{(0)}, \ldots, A^{k-1}r^{(0)}\}.$$

Krylov subspaces are nested and the Arnoldi process (see, e.g., [19]) iteratively computes an orthonormal basis $v^{(1)}, v^{(2)}, \ldots$ for these subspaces. Collecting the vectors into an orthonormal matrix $V^{(k)} = [v^{(1)} \mid \cdots \mid v^{(k)}]$, the Arnoldi process can be summarized by the Arnoldi relation

$$AV^{(k)} = V^{(k+1)}\underline{H}^{(k)}, \quad k = 1, 2, \ldots, \tag{2.1}$$

where $\underline{H}^{(k)} \in \mathbb{C}^{(k+1)\times k}$ collects the coefficients resulting from the orthonormalization process. It has upper Hessenberg structure. Denoting by $H^{(k)}$ the $k \times k$ matrix obtained from $\underline{H}^{(k)}$ by removing the last row, we see that

$$H^{(k)} = (V^{(k)})^* A V^{(k)}.$$

The full orthogonalization method (FOM) is the Krylov subspace method with iterate $x^{(k)}_{\mathtt{fom}}$ characterized variationally via

$$x^{(k)}_{\mathtt{fom}} \in x^{(0)} + \mathcal{K}^{(k)}(A, r^{(0)}), \;\; r^{(k)}_{\mathtt{fom}} = b - Ax^{(k)}_{\mathtt{fom}} \perp \mathcal{K}^{(k)}(A, r^{(0)}),$$

which gives

$$x^{(k)}_{\mathtt{fom}} = x^{(0)} + V^{(k)}(H^{(k)})^{-1}(V^{(k)})^* r^{(0)},$$

provided $H^{(k)}$ is nonsingular. Note that since v_1 is a multiple of $r^{(0)}$ we have

$$(V^{(k)})^* r^{(0)} = \|r^{(0)}\| e_1^k, \tag{2.2}$$

where e_1^k denotes the first canonical unit vector in $\mathbb{C}^k$.

For an arbitrary (nonsingular) matrix A, the matrix $H^{(k)}$ can become singular in which case the k-th FOM iterate does not exist. An important consequence of Lemma 2.1 is, therefore, that such a breakdown of FOM cannot occur if $0 \notin W(A)$. Moreover, $H^{(k)}$ will have no eigenvalues with modulus smaller than the distance of $W(A)$ to 0. On the other hand, if $0 \in W(A)$, even when $H^{(k)}$ is nonsingular, it can become arbitrarily ill-conditioned, which then typically yields large residuals for the corresponding iterates and which is observed in practice as irregular convergence behavior.

We can interpret FOM as the method which for each k builds a reduced model $H^{(k)}$ of dimension k of the original matrix and then obtains its iterate $x^{(k)}_{\mathtt{fom}}$ by lifting the solution of the corresponding reduced system $H^{(k)}\xi_k = (V^{(k)})^* r^{(0)}$ back to the full space as a correction to the initial guess $x^{(0)}$, $x^{(k)}_{\mathtt{fom}} = x^{(0)} + V^{(k)}\xi_k$. This interpretation will serve as a guideline for our development of the "quadratic" FOM method in Section 3.

The generalized minimal residual method (GMRES) is the Krylov subspace method with iterate $x^{(k)}_{\mathtt{gmres}}$ characterized variationally by

$$x^{(k)}_{\mathtt{gmres}} \in x^{(0)} + \mathcal{K}^{(k)}(A, r^{(0)}), \;\; r^{(k)}_{\mathtt{gmres}} = b - Ax^{(k)}_{\mathtt{gmres}} \perp A \cdot \mathcal{K}^{(k)}(A, r^{(0)}).$$

This implies that the residual $b - Ax^{(k)}_{\mathtt{gmres}}$ is smallest in norm among all possible residuals $b - Ax$ with $x \in x^{(0)} + \mathcal{K}^{(k)}(A, r^{(0)})$, i.e., $x^{(k)}_{\mathtt{gmres}}$ solves the least squares problem

$$x^{(k)}_{\mathtt{gmres}} = \operatorname*{argmin}_{x \in x^{(0)} + \mathcal{K}^{(k)}(A, r^{(0)})} \|b - Ax\| = x^{(0)} + \operatorname*{argmin}_{y \in \mathcal{K}^{(k)}(A, r^{(0)})} \|r^{(0)} - Ay\|.$$

To obtain an efficient algorithm it is important to see that this $n \times k$ least squares problem can be reduced to a $(k+1) \times k$ system due to the Arnoldi relation (2.1): we have that $x^{(k)}_{\mathtt{gmres}} = x^{(0)} + V^{(k)}\xi^{(k)}$ where $\xi^{(k)}$ solves

$$\xi^{(k)} = \operatorname*{argmin}_{\xi \in \mathbb{C}^k} \|(V^{(k+1)})^* r^{(0)} - \underline{H}^{(k)}\xi\|, \tag{2.3}$$

where $(V^{(k+1)})^* r^{(0)} = \|r^{(0)}\| e_1^{k+1}$.

In case that $H^{(k)}$ is nonsingular, one can use the normal equation for (2.3) to characterize $\xi_k = (\hat{H}^{(k)})^{-1} e_1^k$, where

$$\hat{H}^{(k)} = H^{(k)} + |h_{k+1,k}|^2((H^{(k)})^{-*}e_k)e_k^*, \text{ where } h_{k+1,k} \text{ is the } (k+1,k) \text{ entry of } \underline{H}^{(k)}.$$

This means that the GMRES approach constructs a reduced model $\hat{H}^{(k)}$ that differs from the FOM model by a matrix of rank 1. The eigenvalues of $\hat{H}^{(k)}$ are called the *harmonic Ritz values* of A w.r.t. $\mathcal{K}^{(k)}(A, r^{(0)})$, i.e., the values μ for which

$$A^{-1}x - \tfrac{1}{\mu}x \perp A\mathcal{K}^{(k)}(A, r^{(0)}) \quad \text{for some } x \in A\mathcal{K}^{(k)}(A, r^{(0)}), x \neq 0.$$

They are the inverses of the Ritz values of A^{-1} w.r.t. the subspace $A\mathcal{K}(A, r^{(0)})$, which implies

$$\mu^{-1} \in W(A^{-1}).$$

With $\rho := \max\{|\omega| : \omega \in W(A^{-1})\}$ denoting the numerical radius of A^{-1} we see that $|\mu| \geq \rho^{-1}$. In this sense, as opposed to FOM, the GMRES approach shields the eigenvalues of the reduced model $\hat{H}^{(k)}$ away from 0. Note that if $H^{(k)}$ is singular, GMRES stagnates, i.e., $x^{(k)}_{\texttt{gmres}} = x^{(k-1)}_{\texttt{gmres}}$.

3. Quadratic numerical range, QFOM, and QGMRES. We now assume that the matrix $A \in \mathbb{C}^{n\times n}$ has a "natural" block decomposition of the form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad \text{with } A_{ij} \in \mathbb{C}^{n_i \times n_j}, i,j = 1,2,\ n_1 + n_2 = n,\ n_1, n_2 \geq 1. \tag{3.1}$$

All vectors x from $\mathbb{C}^n$ are endowed with the same block structure

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad x_i \in \mathbb{C}^{n_i}, i = 1, 2.$$

The definition of the quadratic numerical range goes back to [11], where it was introduced as a tool to localize spectra of block operators in Hilbert spaces; see also [10].

DEFINITION 3.1. *The quadratic numerical range W^2 of A is given as*

$$W^2(A) = \bigcup_{\|x_1\| = \|x_2\| = 1} \operatorname{spec}\left(\begin{bmatrix} x_1^* A_{11} x_1 & x_1^* A_{12} x_2 \\ x_2^* A_{21} x_1 & x_2^* A_{22} x_2 \end{bmatrix} \right).$$

The following basic properties are, for instance, proved in [23].

LEMMA 3.2. *We have*

(i) $W^2(A)$ *is compact,*
(ii) $W^2(A)$ *has at most two connected components,*
(iii) $\operatorname{spec}(A) \subseteq W^2(A) \subseteq W(A)$,
(iv) If $n_1, n_2 \geq 2$, *then* $W(A_{11}), W(A_{22}) \subseteq W^2(A)$.

The quadratic numerical range is an established tool in operator theory where typically infinite-dimensional operators are considered. If $W^2(A)$ has two components, then all $\lambda \in \mathbb{C}$ that do not belong to any of the two components but to $W(A)$ cannot be eigenvalues, thus establishing the existence of spectral gaps. Such gaps are important in many applications, like for example for the electrical or optical conductivity in materials. Usually, the separation of two components of $W^2(A)$ is shown using analytical approaches. Despite some efforts for special finite-dimensional operators in [15, 16], efficient numerical approaches to accurately approximate $W^2(A)$ are still missing. The illustrations in the book [23], for example, use stochastic sampling, which works only for very small dimensions.

The following counterpart of Lemma 2.1 holds.

LEMMA 3.3. *Let $A \in \mathbb{C}^{n\times n}$ have block structure* (3.1) *and assume that $V_1 \in \mathbb{C}^{n_1 \times m_1}$, $V_2 \in \mathbb{C}^{n_2 \times m_2}$ with $m_i \leq n_i, i = 1, 2$ have orthonormal columns. Put $V = \left[\begin{smallmatrix} V_1 & 0 \\ 0 & V_2 \end{smallmatrix}\right] \in \mathbb{C}^{n\times m}$ with $m = m_1 + m_2$. Then*

$$W^2(V^*AV) \subseteq W^2(A), \qquad \textit{where} \qquad V^*AV = \begin{bmatrix} V_1^* A_{11} V_1 & V_1^* A_{12} V_2 \\ V_2^* A_{21} V_1 & V_2^* A_{22} V_2 \end{bmatrix} \in \mathbb{C}^{m\times m}.$$

Proof. Let $y_i \in \mathbb{C}^{m_i}$ for $i = 1, 2$ with $\|y_i\| = 1$. Then $x_i := V_i y_i$ satisfies $\|x_i\| = 1$, $i = 1, 2$, and since

$$\begin{bmatrix} (y_1)^* V_1^* A_{11} V_1 y_1 & (y_1)^* V_1^* A_{12} V_2 y_2 \\ (y_2)^* V_2^* A_{21} V_1 y_1 & (y_2)^* V_2^* A_{22} V_2 y_2 \end{bmatrix} = \begin{bmatrix} x_1^* A_{11} x_1 & x_1^* A_{12} x_2 \\ x_2^* A_{21} x_1 & x_2^* A_{22} x_2 \end{bmatrix}$$

we obtain $W^2(V^*AV) \subseteq W^2(A)$. □

Our approach is now to build a Krylov subspace type method where, as opposed to FOM, the iterates are obtained by inverting a reduced model of A whose quadratic numerical range is contained in that of A. In this manner, if $0 \notin W^2(A)$ with $\delta = \min\{|\mu| : \mu \in W^2(A)\}$ denoting the distance of 0 to $W^2(A)$, no eigenvalue of the reduced model will have modulus smaller than δ. In cases where $0 \in W(A)$ and $0 \notin W^2(A)$ this bears the potential of obtaining smoother and faster convergence than with FOM and, as it will turn out experimentally, also faster than with GMRES.

We project the Krylov subspace $\mathcal{K}^{(k)}(A, r^{(0)})$ onto its first n_1 and last n_2 components, respectively, denoted $\mathcal{K}_1^{(k)}(A, r^{(0)}) \subseteq \mathbb{C}^{n_1}$ and $\mathcal{K}_2^{(k)}(A, r^{(0)}) \subseteq \mathbb{C}^{n_2}$. Clearly,

$$\mathcal{K}^{(k)}(A, r^{(0)}) \subseteq \mathcal{K}_1^{(k)}(A, r^{(0)}) \times \mathcal{K}_2^{(k)}(A, r^{(0)}) =: \mathcal{K}_\times^{(k)}(A, r^{(0)}),$$

where × denotes the Cartesian product, and $\dim \mathcal{K}^{(k)}(A, r^{(0)}) \leq \dim \mathcal{K}_\times^{(k)}(A, r^{(0)}) =: d_\times^{(k)} \leq 2k$. Note that the dimension $d_i^{(k)}$ of either $\mathcal{K}_i^{(k)}(A, r^{(0)})$ may be less than k and $d_\times^{(k)} = d_1^{(k)} + d_2^{(k)}$.

We can obtain an orthonormal basis for each of the $\mathcal{K}_i^{(k)}(A, r^{(0)})$ as the columns of the matrix $V_i^{(k)}$ that arises from the QR-decomposition of the respective block of the matrix $V^{(k)}$ from the Arnoldi process, i.e.,

$$V^{(k)} = \begin{bmatrix} V_1^{(k)} R_1^{(k)} \\ V_2^{(k)} R_2^{(k)} \end{bmatrix}, \quad V_i^{(k)} \in \mathbb{C}^{n_i \times d_i^{(k)}} \text{ orthonorm.}, \; R_i^{(k)} \in \mathbb{C}^{d_i^{(k)} \times k} \text{ upper triang.} \tag{3.2}$$

Note that with this definition of $V_i^{(k)}$ we have the useful property that $V_i^{(k+1)}$ arises from $V_i^{(k)}$ by the addition of a new last column, just in the way $V^{(k+1)}$ arises from $V^{(k)}$, with the exception that the new last column could be empty, i.e., there is no new last column, when the last column of the i-th block in $V^{(k)}$ is linearly dependent of the other columns. Similarly, $R_i^{(k+1)}$ arises from $R_i^{(k)}$ by adding a new last column and a new last row (if it is not empty).

We now introduce variational characterizations based on the space $\mathcal{K}_\times^{(k)}(A, r^{(0)})$.

3.1. QFOM.

Quadratic FOM imposes a Galerkin condition using $\mathcal{K}_\times^{(k)}(A, r^{(0)})$.

DEFINITION 3.4. *The k-th* quadratic FOM ("QFOM") iterate $x_{\texttt{qfom}}^{(k)}$ *is defined variationally through*

$$x_{\texttt{qfom}}^{(k)} \in x^{(0)} + \mathcal{K}_\times^{(k)}(A, r^{(0)}), \;\; b - Ax_{\texttt{qfom}}^{(k)} \perp \mathcal{K}_{(\times)}^{(k)}(A, r^{(0)}). \tag{3.3}$$

The columns of the matrix

$$V_\times^{(k)} = \begin{bmatrix} V_1^{(k)} & 0 \\ 0 & V_2^{(k)} \end{bmatrix}$$

form an orthonormal basis of $\mathcal{K}_\times^{(k)}(A, r^{(0)})$. Defining the reduced model $H_\times^{(k)}$ of A as

$$H_\times^{(k)} = (V_\times^{(k)})^* A V_\times^{(k)} = \begin{bmatrix} (V_1^{(k)})^* A_{11} V_1^{(k)} & (V_1^{(k)})^* A_{12} V_2^{(k)} \\ (V_2^{(k)})^* A_{21} V_1^{(k)} & (V_2^{(k)})^* A_{22} V_2^{(k)} \end{bmatrix} \tag{3.4}$$

we see that if $H_\times^{(k)}$ is nonsingular, the QFOM iterate $x_{\mathtt{qfom}}^{(k)}$ according to Definition 3.4 exists and can be represented as

$$x_{\mathtt{qfom}}^{(k)} = x^{(0)} + V_\times^{(k)}(H_\times^{(k)})^{-1}(V_\times^{(k)})^* r^{(0)}.$$

Instead of (2.2) we now have

$$(V_\times^{(k)})^* r^{(0)} = \begin{bmatrix} \|r_1^{(0)}\| e_1^{d_1^{(k)}} \\ \|r_2^{(0)}\| e_1^{d_2^{(k)}} \end{bmatrix}, \text{ where } r^{(0)} = \begin{bmatrix} r_1^{(0)} \\ r_2^{(0)} \end{bmatrix}. \tag{3.5}$$

If $H_\times^{(k)}$ is singular, the k-th QFOM iterate does not exist. We will show in Section 4 that computing $x_{\mathtt{qfom}}^{(k)}$ costs k matrix-vector multiplications with A plus additional arithmetic operations of order $\mathcal{O}(k^3)$. The cost is therefore the same as for standard FOM in terms of matrix-vector multiplications, and the additional cost is also of the same order (though with a larger constant).

3.2. Properties of QFOM. The following theorem summarizes some basic properties of QFOM. Recall that the *grade* of a vector v with respect to a square matrix A is the first index $g(v)$ for which $\mathcal{K}^{(g(v))}(A,v) = \mathcal{K}^{(g(v)+1)}(A,v)$. We know (see, e.g, [19]) that then $\mathcal{K}^{(g(v))}(A,v) = \mathcal{K}^{(g(v)+i)}(A,v)$ for all $i \geq 0$ and that $A^{-1}v \in \mathcal{K}^{(g(v))}(A,v)$, provided A is nonsingular.

THEOREM 3.5. *Let A be nonsingular. Then,*

(i) [Finite termination] there exists an index $k_{\max} \leq g(r^{(0)})$ such that $A^{-1}r^{(0)} \in \mathcal{K}_\times^{(k_{\max})}(A, r^{(0)})$, and if $H_\times^{(k_{\max})}$ is nonsingular, $x_{\mathtt{qfom}}^{(k_{\max})}$ exists and $x_{\mathtt{qfom}}^{(k_{\max})} = A^{-1}b$;

(ii) [Quadratic numerical range property] the inclusion $W^2(H_\times^{(k)}) \subseteq W^2(A)$ holds for $k = 1, \ldots, k_{\max}$, where the 2×2 block structure of $H_\times^{(k)}$ is given in (3.4);

(iii) [Existence] if $0 \notin W^2(A)$, then $x_{\mathtt{qfom}}^{(k)}$ exists for $k = 1, \ldots, k_{\max}$, i.e., $H_k^\times$ is nonsingular for all $k = 1, \ldots, k_{\max}$.

Proof. To show (i), let g be the grade of $r^{(0)}$ w.r.t. A and let $k_{\max} \leq g$ be the smallest index k for which $\mathcal{K}^{(g)}(A, r^{(0)}) \subseteq \mathcal{K}_\times^{(k)}(A, r^{(0)})$. Since A is nonsingular, there exists $y^* \in \mathcal{K}_\times^{(k_{\max})}(A, r^{(0)})$ with $Ay^* = r^{(0)}$, i.e., $y^* = A^{-1}r^{(0)}$. As a consequence, $x^* = A^{-1}b = x^{(0)} + y^* \in x^{(0)} + \mathcal{K}_\times^{(k_{\max})}(A, r^{(0)})$ satisfies the variational characterization (3.3) from Definition 3.4 just as $x_{\mathtt{qfom}}^{(k_{\max})}$ does. If $H_\times^{(k_{\max})}$ is nonsingular there is exactly one vector from $x^{(0)} + \mathcal{K}_\times^{(k_{\max})}(A, r^{(0)})$ which satisfies (3.3) which gives $x_{\mathtt{qfom}}^{(k_{\max})} = x^*$.

Part (ii) follows directly from Lemma 3.3. Finally, part (iii) is an immediate consequence of part (ii) and the spectral enclosure property stated as Lemma 3.2 (iii). □

More far-reaching results seem to be difficult to obtain. In particular, the absence of a polynomial interpolation property—which we discuss in the sequel—makes it impossible to follow established concepts from standard Krylov subspace theory.

The FOM iterates satisfy a polynomial interpolation property: We know that $(H^{(k)})^{-1} = q(H^{(k)})$, where q is the polynomial of degree at most $k-1$ that interpolates the function $z \to z^{-1}$ on the eigenvalues in the Hermite sense, i.e., up to the $j-1$st derivative if the multiplicity of the eigenvalue in the minimal polynomial is j; see [7]. We have that

$$V^{(k)}(H^{(k)})^{-1}(V^{(k)})^* r^{(0)} = V^{(k)} q(H^{(k)})(V^{(k)})^* r^{(0)} = q(A) r^{(0)},$$

where the last, important equality holds because $V^{(k)}(V^{(k)})^*$ represents the orthogonal projector on $\mathcal{K}_k(A, r^{(0)})$, thus implying that for all powers $j = 0, \ldots, k-1$, we have

$V^{(k)}(H^{(k)})^j(V^{(k)})^*r^{(0)} = V^{(k)}((V^{(k)})^*AV^{(k)})^j(V^{(k)})^*r^{(0)} = A^jr^{(0)}$; see also the moment matching property of FOM as treated in, e.g., [12]. As a consequence,

$$x^{(k)}_{\texttt{fom}} = x^{(0)} + q(A)r^{(0)}. \tag{3.6}$$

Since $\hat{H}^{(k)}$ differs from $H^{(k)}$ only in its last column, the same argument as above shows that an analogue of (3.6) holds for the GMRES iterates, where now q interpolates on the spectrum of $\hat{H}^{(k)}$. This interpolation property is very helpful in the analysis of the FOM and GMRES method, but there is no analogue for QFOM. Indeed, while we can express $(H^{(k)}_\times)^{-1}$ as a polynomial q of degree at most $d^{(k)}_1 + d^{(k)}_2 - 1 \le 2k-1$ in $H^{(k)}_\times$, the matrix $V^{(k)}_\times(V^{(k)}_\times)^*$ is an orthogonal projector on $K^{(k)}_\times(A, r^{(0)})$, which contains $K^{(k)}(A, r^{(0)})$, but not necessarily the higher powers $A^ir^{(0)}$ for $i \ge k$. Therefore, we cannot conclude that $(V^{(k)}_\times)(H^{(k)}_\times)^i(V^{(k)}_\times)^*r^{(0)} = (V^{(k)}_\times)((V^{(k)}_\times)^*AV^{(k)}_\times)^i(V^{(k)}_\times)^*r^{(0)}$ would be equal to $A^ir^{(0)}$ for $i \ge k$, and therefore, since the degree of the polynomial q is likely to be larger than $k-1$ we do not get $V^{(k)}_\times q(H^{(k)}_\times)(V^{(k)}_\times)^*r^{(0)} = q(A)r^{(0)}$.

To finish this section, we look at the very extreme case in which $W^2(A)$ consists of just one or two points. We show that in this case QFOM obtains the solution after just one iteration in a larger number of cases than standard FOM or GMRES does. So assume $W^2(A) = \{\lambda_1, \lambda_2\}$, where $\lambda_1 = \lambda_2$ is allowed.

LEMMA 3.6. *Let $n_1, n_2 \ge 2$. $W^2(A) = \{\lambda_1, \lambda_2\}$ if and only if*

$$A = \begin{bmatrix} \lambda_1 I & A_{12} \\ A_{21} & \lambda_2 I \end{bmatrix}, \quad \textit{where } A_{12} = 0 \textit{ or } A_{21} = 0, \tag{3.7}$$

up to a permutation of λ_1, λ_2 on the diagonal.

Proof. For $x_i \in \mathbb{C}^{n_i}$, $\|x_i\| = 1, i = 1, 2$ denote

$$\alpha = x_1^*A_{11}x_1, \beta = x_1^*A_{12}x_2, \gamma = x_2^*A_{21}x_1, \delta = x_2^*A_{22}x_2.$$

Then $\lambda \in W^2(A)$ iff

$$(\lambda - \alpha)(\lambda - \delta) - \beta\gamma = 0 \tag{3.8}$$

for $\alpha, \beta, \gamma, \delta$ associated with such x_1, x_2. Now, if A is of the form (3.7), then $\beta\gamma = 0$, $\alpha = \lambda_1$, and $\delta = \lambda_2$, which immediately gives that (3.7) is sufficient to get $W^2(A) = \{\lambda_1, \lambda_2\}$.

To prove necessity, assume $W^2(A) = \{\lambda_1, \lambda_2\}$. Since $W(A_{ii}) \subseteq W^2(A)$ for $i = 1, 2$ by Lemma 3.2 (iv) and since the numerical range is convex, this implies $W(A_{11}) = \{\mu_1\}$, $W(A_{22}) = \{\mu_2\}$ with $\mu_1, \mu_2 \in \{\lambda_1, \lambda_2\}$. Consequently $A_{11} = \mu_1 I, A_{22} = \mu_2 I$. For a proof by contradiction assume now that both A_{12} and A_{21} are nonzero. Then there exist normalized vectors x_1, x_2, y_1, y_2 such that $x_1^*A_{12}x_2 \ne 0$ and $y_2^*A_{21}y_1 \ne 0$. For $\epsilon \in \mathbb{R}$, consider $z_1 = x_1 + \epsilon y_1, z_2 = x_2 + \epsilon y_2$. Then $z_1^*A_{12}z_2 \ne 0$ for $\epsilon \ne 0$ small enough and

$$z_2^*A_{21}z_1 = x_2^*A_{21}x_1 + \epsilon(x_2^*A_{21}y_1 + y_2^*A_{21}x_1) + \epsilon^2 y_2^*A_{21}y_1.$$

This quadratic function in ϵ is nonzero for sufficiently small $\epsilon \ne 0$. Thus, for $\epsilon \ne 0$ sufficiently small, taking the normalized versions of z_1, z_2 we get that the corresponding β and γ are both nonzero. Consequently, the expression

$$(\lambda - \mu_1)(\lambda - \mu_2) - \beta\gamma$$

is nonzero for $\lambda = \mu_1 \in W^2(A)$, but zero at the same time by (3.8). Thus, at least one of the matrices A_{12}, A_{21} is zero. It follows that $W^2(A) = \{\mu_1, \mu_2\}$ and consequently $\mu_1 = \lambda_1$ and $\mu_2 = \lambda_2$ up to a permutation of λ_1, λ_2. □

With these preparations we obtain the following result.

THEOREM 3.7. *Assume that $n_1, n_2 \geq 2$ and $0 \notin W^2(A) = \{\lambda_1, \lambda_2\}$. Consider the linear system*

$$Ax = b.$$

Without loss of generality we assume that iterations start with the initial guess $x^{(0)} = 0$. We also denote by $x^* = A^{-1}b$ the solution of the system. Then

(i) $x^{(1)}_{\texttt{fom}} = x^*$ if b is an eigenvector of A. In all other cases, $x^{(2)}_{\texttt{fom}} = x^*$.

(ii) $x^{(1)}_{\texttt{qfom}} = x^*$ if $A_{12}b_2$ is collinear to b_1 (or 0). In all other cases, $x^{(2)}_{\texttt{qfom}} = x^*$.

Proof. By Lemma 3.6 we know that A has the form

$$A = \begin{bmatrix} \lambda_1 I & A_{12} \\ 0 & \lambda_2 I \end{bmatrix} \quad \text{or} \quad A = \begin{bmatrix} \lambda_1 I & 0 \\ A_{21} & \lambda_2 I \end{bmatrix},$$

and we focus on the first case. The second case can be treated in a completely analogous manner. We first note that if $\lambda_1 \neq \lambda_2$, the eigenvectors to the eigenvalue λ_1 are of the form $\left[\begin{smallmatrix} x_1 \\ 0 \end{smallmatrix}\right]$ and the eigenvectors to the eigenvalue λ_2 are given by $\left[\begin{smallmatrix} (\lambda_2-\lambda_1)^{-1}A_{12}x_2 \\ x_2 \end{smallmatrix}\right]$. If $\lambda_1 = \lambda_2$, all vectors of the form $\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]$ with $A_{12}x_2 = 0$ are eigenvectors. The theorem, thus, asserts that the situations where FOM finds the solution in the first iteration are a true subset of the situations in which QFOM obtains the solution in its first iteration.

To proceed, we observe that the minimal polynomial of A is $p(z) = (z - \lambda_1)(z - \lambda_2)$ in all cases except for the case where $\lambda_1 = \lambda_2$ and $A_{12} = 0$, i.e., when $A = \lambda_1 I$ with minimal polynomial $p(z) = (z - \lambda_1)$. Since $x^{(1)}_{\texttt{fom}} \in \mathcal{K}^{(1)}(A, b)$, which is spanned by b, FOM obtains the solution x^* in the first iteration exactly in the case where b is an eigenvector of A. If b is not an eigenvector of A, then the minimal polynomial is $p(z) = (z - \lambda_1)(z - \lambda_2)$ so that the grade of b is 2, and FOM obtains the solution x^* in its second iteration.

If $b_1 \neq 0$ and $b_2 \neq 0$, the first iteration of QFOM obtains $x^{(1)}_{\texttt{qfom}}$ as

$$\begin{aligned}
x^{(1)}_{\texttt{qfom}} &= \begin{bmatrix} \frac{1}{\|b_1\|}b_1 & 0 \\ 0 & \frac{1}{\|b_2\|}b_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & \frac{1}{\|b_1\|\,\|b_2\|}b_1^* A_{12} b_2 \\ 0 & \lambda_2 \end{bmatrix}^{-1} \begin{bmatrix} \|b_1\| \\ \|b_2\| \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\|b_1\|}b_1 & 0 \\ 0 & \frac{1}{\|b_2\|}b_2 \end{bmatrix} \begin{bmatrix} \frac{1}{\lambda_1} & -\frac{1}{\lambda_1\lambda_2}\frac{1}{\|b_1\|\,\|b_2\|}b_1^* A_{12} b_2 \\ 0 & \frac{1}{\lambda_2} \end{bmatrix} \begin{bmatrix} \|b_1\| \\ \|b_2\| \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\lambda_1}(b_1 - \frac{1}{\lambda_2}\frac{1}{\|b_1\|^2}b_1 b_1^* A_{12} b_2) \\ \frac{1}{\lambda_2}b_2 \end{bmatrix},
\end{aligned}$$

which is equal to the solution

$$x^* = \begin{bmatrix} \frac{1}{\lambda_1}(b_1 - \frac{1}{\lambda_2}A_{12}b_2) \\ \frac{1}{\lambda_2}b_2 \end{bmatrix}$$

exactly when the projector $\frac{1}{\|b_1\|^2}b_1 b_1^*$ acts as the identity on $A_{12}b_2$, i.e., when $A_{12}b_2$ is zero or collinear to b_1. A similar observation holds if $b_1 = 0$ or $b_2 = 0$. In all other cases, by Theorem 3.5 we have $x^{(2)}_{\texttt{qfom}} = x^*$ since the grade of b then equals 2. □

3.3. QGMRES and QQGMRES.

In principle, we can proceed in a manner similar to QFOM to derive a "quadratic" GMRES method. Variationally, its iterates $x^{(k)}_{\texttt{qgmr}}$ would be characterized by

$$x^{(k)}_{\texttt{qgmr}} \in x^{(0)} + \mathcal{K}^{(k)}_{\times}(A, r^{(0)}), \;\; b - Ax^{(k)}_{\texttt{qgmr}} \perp A\mathcal{K}^{(k)}_{\times}(A, r^{(0)}), \tag{3.9}$$

which is equivalent to minimizing the norm of the residual $\|b - Ax\|$ for the vector $x \in x^{(0)} + K_\times^{(k)}(A, r^{(0)})$. Thus, as for standard GMRES, we can compute the k-th iterate $x_{\mathtt{qgmr}}^{(k)}$ as $x^{(0)} + V_\times^{(k)}\eta_k$, where η_k solves the least squares problem

$$\eta_k = \operatorname*{argmin}_{\eta \in \mathbb{C}^{d_\times^{(k)}}} \|r^{(0)} - AV_\times^{(k)}\eta\|. \tag{3.10}$$

However, as opposed to standard GMRES, it is not possible to recast this $n \times d_\times^{(k)}$ least squares problem reducing the first dimension, since an analogue to the Arnoldi relation (2.1) does not hold for the product spaces $\mathcal{K}_\times^{(k)}(A, r^{(0)})$. In particular, for $x^{(k)} \in x^{(0)} + \mathcal{K}_\times^{(k)}(A, r^{(0)})$, the residual $r^{(k)} = r^{(0)} - Ax^{(k)}$ need not be contained in $\mathcal{K}_\times^{(k+1)}(A, r^{(0)})$. This fact prevents approaches based on the variational characterization (3.9) to be realized with cost depending exclusively on k and not on n.

As an alternative, we thus suggest an approach similar to truncated GMRES; see, e.g., [19]. We project the $n \times d_\times^{(k)}$ least squares problem (3.10) onto a $d_\times^{(k+1)} \times d_\times^{(k)}$ least squares problem by minimizing, instead of the whole residual $\|b - Ax^{(k)}\|$, only its orthogonal projection on $\mathcal{K}_\times^{(k+1)}(A, r^{(0)})$.

DEFINITION 3.8. *The k-th* quadratic quasi GMRES ("QQGMRES") *iterate $x_{\mathtt{qqgmr}}^{(k)}$ is the solution of the least squares problem*

$$x_{\mathtt{qqgmr}}^{(k)} = \operatorname*{argmin}_{x \in x^{(0)} + \mathcal{K}_\times^{(k)}(A,b)} \|(V_\times^{(k+1)})^*(b - Ax)\|.$$

Computationally, we have that $x_{\mathtt{qqgmr}}^{(k)} = x^{(0)} + V_\times^{(k)}\zeta_k$, where ζ_k solves the $d_\times^{(k+1)} \times d_\times^{(k)}$ least squares problem

$$\zeta_k = \operatorname*{argmin}_{\zeta \in \mathbb{C}^{d_\times^{(k)}}} \|(V_\times^{(k+1)})^* r^{(0)} - \underline{H}_\times^{(k)}\zeta\|, \tag{3.11}$$

where

$$\underline{H}_\times^{(k)} = (V_\times^{(k+1)})^* A V_\times^{(k)} = \begin{bmatrix} (V_1^{(k+1)})^* A_{11} V_1^{(k)} & (V_1^{(k+1)})^* A_{12} V_2^{(k)} \\ (V_2^{(k+1)})^* A_{21} V_1^{(k)} & (V_2^{(k+1)})^* A_{22} V_2^{(k)} \end{bmatrix} \tag{3.12}$$

and where the structure of $(V_\times^{(k+1)})^* r^{(0)}$ is given in (3.5). In general, $\underline{H}_\times^{(k)}$ need not have full (column) rank, in which case (3.11) will have more than one solution. The QQGMRES iterate is then not uniquely defined unless we impose additional conditions on ζ_k from (3.11) like having minimal norm, for example. Proposition 3.9 below will show that this more involved situation does not occur if $0 \notin W^2(A)$.

3.4. Properties of QGMRES and QQGMRES.

As for QFOM, there is no polynomial interpolation property for QGMRES nor for QQGMRES. We can again present only simple first elements of an analysis.

As solutions to least squares problems, the iterates $x_{\mathtt{qgmr}}^{(k)}$ and $x_{\mathtt{qqgmr}}^{(k)}$ are always defined. They are uniquely defined in case of QGMRES, since $AV_\times^{(k)}$ has full rank since $V_\times^{(k)}$ has full rank. For QQGMRES we prove the following statement.

PROPOSITION 3.9. *The matrix $\underline{H}_\times^{(k)}$ from* (3.12) *has full rank if* $0 \notin W^2(A)$.

Proof. The matrix $\underline{H}_\times^{(k)}$ is obtained from $H_\times^{(k)}$ by complementing it with two rows, one after each block, and $H_\times^{(k)}$ is nonsingular by Theorem 3.5 (iii). Thus, $\underline{H}_\times^{(k)}$ has full rank $d_\times^{(k)} = d_1^{(k)} + d_2^{(k)}$. □

QGMRES and QQGMRES also both have a finite termination property.

PROPOSITION 3.10. *Let $k_{\max} \leq g(r^{(0)})$ be as in the proof of Theorem 3.5. Then $x_{\texttt{qgmr}}^{(k_{\max})} = A^{-1}b$. Provided $\underline{H}_{\times}^{(k_{\max})}$ has full rank, we also have $x_{\texttt{qqgmr}}^{(k_{\max})} = A^{-1}b$.*

Proof. As in the proof of Theorem 3.5, we have that $x^* = A^{-1}b = x^{(0)} + y^*$ with $y^* = A^{-1}r^{(0)}$ being contained in $\mathcal{K}_{\times}^{(k_{\max})}(A, r^{(0)})$. So x^* satisfies the variational characterization (3.9) with residual norm 0, and as such it is unique. This implies that x^* is identical to the QGMRES iterate $x_{\texttt{qgmr}}^{(k_{\max})}$. For QQGMRES, we write $y^* \in \mathcal{K}_{\times}^{(k_{\max})}(A, r^{(0)})$ as $y^* = V_{\times}^{(k)}\zeta$. This ζ is a solution of the least squares problem (3.11), yielding the minimal value 0 for the residual norm. If $\underline{H}_{\times}^{(k_{\max})}$ has full rank, the solution of the least squares problem (3.11) is unique. And since the QQGMRES iterate $x_{\texttt{qqgmr}}^{(k_{\max})}$ is obtained by solving this least squares problem, it is equal to x^*. □

Trivially, QGMRES computes iterates $x_{\texttt{qgmr}}^{(k)}$ whose residuals $r_{\texttt{qgmr}}^{(k)}$ are smaller in norm than $r_{\texttt{gmres}}^{(k)}$, i.e., the residual of the k-th iterate $x_{\texttt{gmres}}^{(k)}$ of standard GMRES, since QGMRES minimizes the residual norm over a larger subspace. Moreover, since QQGMRES minimizes over the same subspace as QGMRES, but minimizes the norm of the projection of the residual rather than the norm of the residual itself, we also have that $\|r_{\texttt{qgmr}}^{(k)}\| \leq \|r_{\texttt{qqgmr}}^{(k)}\|$. Finally, note that we cannot expect the relation $\|r_{\texttt{qqgmr}}^{(k)}\| \leq \|r_{\texttt{gmres}}^{(k)}\|$ to hold in general.

The residuals of the standard FOM and GMRES iterates satisfy (see, e.g., [19])

$$\|r_{\texttt{fom}}^{(k)}\|_2^2 = \|r_{\texttt{gmres}}^{(k)}\|_2^2 \,/\, \Big(1 - \frac{\|r_{\texttt{gmres}}^{(k)}\|_2^2}{\|r_{\texttt{gmres}}^{(k-1)}\|_2^2}\Big), \tag{3.13}$$

which immediately explains that peaks in FOM are related to stagnation in GMRES. Relation (3.13) genuinely relies on the fact that the approximation subspaces are Krylov subspaces. It, thus, does not carry over to a similar relation between the QFOM and QGMRES residual norms, and also not to QQGMRES residual norms.

4. Algorithmic aspects.

An important practical question is how one can compute $V_{\times}^{(k)}$ and $H_{\times}^{(k)}$ efficiently and in a stable manner. Interestingly, for the special case where $A_{21} = I$ and $A_{22} = 0$, which arises in the linearization of quadratic eigenvalue problems, this question has been treated in many papers, starting with the second-order Arnoldi method from [1]. Recently, the *two-level orthogonal Arnoldi method* has emerged as a cost-efficient and at the same time stable algorithm; see [9, 13, 14]. In the following, we describe how these approaches generalize to arbitrary 2×2 block matrices. We note that generalizing the stability analysis of [9, 13, 14] is not straightforward, and a detailed analysis is beyond the scope of this paper.

The main idea is that we refrain from directly computing the orthogonal Arnoldi basis $V^{(k)}$ from (2.1), but rather compute/update the orthonormal bases $V_1^{(k)}, V_2^{(k)}$ of its block components while at the same time updating $H_{\times}^{(k)}$.

Assume that no breakdown occurs and no deflation is necessary. Then we have (see (3.2))

$$V^{(k)} = \begin{bmatrix} V_1^{(k)} R_1^{(k)} \\ V_2^{(k)} R_2^{(k)} \end{bmatrix},$$

where the $V_i^{(k)}$ have k orthonormal columns, and the $R_i^{(k)} \in \mathbb{C}^{k \times k}$ are upper triangular. Since the columns of $V^{(k)}$ are orthonormal, too, this implies

$$(R_1^{(k)})^* R_1^{(k)} + (R_2^{(k)})^* R_2^{(k)} = (V^{(k)})^* V^{(k)} = I, \tag{4.1}$$

showing that the matrix $\begin{bmatrix} R_1^{(k)} \\ R_2^{(k)} \end{bmatrix} \in \mathbb{C}^{2k\times k}$ also has orthonormal columns. Writing the Arnoldi relation (2.1) in terms of the block components gives

$$
\begin{aligned}
A_{11}V_1^{(k)}R_1^{(k)} + A_{12}V_2^{(k)}R_2^{(k)} &= V_1^{(k+1)}R_1^{(k+1)}\underline{H}^{(k)} =: V_1^{(k+1)}\underline{H}_1^{(k)}, \\
A_{21}V_1^{(k)}R_1^{(k)} + A_{22}V_2^{(k)}R_2^{(k)} &= V_2^{(k+1)}R_2^{(k+1)}\underline{H}^{(k)} =: V_2^{(k+1)}\underline{H}_2^{(k)},
\end{aligned} \tag{4.2}
$$

where the matrices

$$
\underline{H}_i^{(k)} := R_i^{(k+1)}\underline{H}^{(k)} \in \mathbb{C}^{(k+1)\times k}, \qquad i = 1, 2,
$$

are upper Hessenberg.

The relation (4.2) reveals that $V_i^{(k+1)}$ can be obtained as an update of $V_i^{(k)}$ by adding a new last column, and $\underline{H}_i^{(k)}$ as an update of $\underline{H}_i^{(k-1)}$ by adding a new last column and a new last row. Thus, the new column of $V_i^{(k+1)}$ arises from the orthonormalization of the last column of $A_{i1}V_1^{(k)}R_1^{(k)} + A_{i2}V_2^{(k)}R_2^{(k)}$ against all columns of $V_i^{(k)}$ and it is nonzero. The upper Hessenberg matrix $\underline{H}_1^{(k)}$ is obtained from $\underline{H}_1^{(k-1)}$ by first adding a new last row of zeros and then adding a new last column holding the coefficients from the orthonormalization. To obtain a viable computational scheme, it remains to show that $R_i^{(k+1)}$ as well as $\underline{H}^{(k)}$ (which we need to get the QFOM or QGMRES iterates) can also be obtained from these quantities. We do so by establishing how to get them as updates from $H^{(k-1)}$ and $R_i^{(k)}$. In the very first step we have

$$
R_i^{(1)} = \|b_i\|, \qquad V_i^{(1)} = b_i/\|b_i\|, \qquad i = 1, 2,
$$

unless $b_i = 0$ in which case we let the corresponding $R_i^{(1)}$ be zero and let $V_i^{(1)}$ be a random unitary vector.

For $k > 1$, we write

$$
R_i^{(k+1)} = \begin{bmatrix} R_i^{(k)} & r_i^{(k+1)} \\ 0 & \rho_i^{(k+1)} \end{bmatrix} \quad \text{and} \quad \underline{H}^{(k)} = \begin{bmatrix} \underline{H}^{(k-1)} & h^{(k)} \\ 0 & \eta^{(k)} \end{bmatrix},
$$

where $R_i^{(k)}$ and $\underline{H}^{(k-1)}$ are known, and the remaining quantities are to be determined. Since $\underline{H}_i^{(k)}$ equals

$$
R_i^{(k+1)}\underline{H}^{(k)} = \begin{bmatrix} R_i^{(k)}\underline{H}^{(k-1)} & R_i^{(k)}h^{(k)} + \eta_i^{(k)}r_i^{(k+1)} \\ 0 & \eta^{(k)}\rho_i^{(k+1)} \end{bmatrix} = \begin{bmatrix} \underline{H}_i^{(k-1)} & h_i^{(k)} \\ 0 & \eta_i^{(k)} \end{bmatrix}, \tag{4.3}
$$

it follows, using (4.1), that

$$
\begin{aligned}
[(R_1^{(k)})^*\ 0]\underline{H}_1^{(k)} + [(R_2^{(k)})^*\ 0]\underline{H}_2^{(k)} &= \big((R_1^{(k)})^*[R_1^{(k)}\ r_1^{(k+1)}] + (R_2^{(k)})^*[R_2^{(k)}\ r_2^{(k+1)}]\big)\underline{H}^{(k)} \\
&= [I\ 0]\underline{H}^{(k)} \ = \ H^{(k)}.
\end{aligned}
$$

Hence, we see that

$$
h^{(k)} = (R_1^{(k)})^*h_1^{(k)} + (R_2^{(k)})^*h_2^{(k)},
$$

which allows for the computation of $h^{(k)}$ from known quantities. Once $h^{(k)}$ is known, (4.3) can be used to compute

$$
\widetilde{r}_i^{(k+1)} = \eta^{(k)}r_i^{(k+1)} = h_i^{(k)} - R_i^{(k)}h^{(k)},
$$

at which point $\eta^{(k)}$ and the $\rho_i^{(k)}$ are the only remaining quantities to be determined. Letting $\eta^{(k)}$ be real valued (and nonnegative) allows its computation in at least two different ways. The first is to consider the bottom right entry of (4.1), which gives

$$\begin{aligned}(\eta^{(k)})^2 &= \|\eta^{(k)} r_1^{(k+1)}\|^2 + |\eta^{(k)} \rho_1^{(k+1)}|^2 + \|\eta^{(k)} r_2^{(k+1)}\|^2 + |\eta^{(k)} \rho_2^{(k+1)}|^2 \\ &= \|\widetilde{r}_1^{(k+1)}\|^2 + |\eta_1^{(k)}|^2 + \|\widetilde{r}_2^{(k+1)}\|^2 + |\eta_2^{(k)}|^2.\end{aligned}$$

The second possibility is to determine $\eta^{(k)}$ from the $(k+1, k+1)$ entry of the equality $(\underline{H}^{(k)})^* \underline{H}^{(k)} = (\underline{H}_1^{(k)})^* \underline{H}_1^{(k)} + (\underline{H}_2^{(k)})^* \underline{H}_2^{(k)}$, which results in

$$(\eta^{(k)})^2 + \|h^{(k)}\|^2 = \|h_1^{(k)}\|^2 + |\eta_1^{(k)}|^2 + \|h_2^{(k)}\|^2 + |\eta_2^{(k)}|^2,$$

using (4.1). The first method may be preferred, since it guarantees that the computed $(\eta^{(k)})^2$ is nonnegative, even with roundoff errors. Once $\eta^{(k)}$ has been determined, we get $\rho_i^{(k)}$ as $\rho_i^{(k)} = \eta_i^{(k)}/\eta^{(k)}$ from (4.3). Putting everything together yields the following proposition.

PROPOSITION 4.1. *In iteration k, the quantities $V_i^{(k+1)}$, $R_i^{(k+1)}$, and $\underline{H}_i^{(k)}$ as well as $\underline{H}^{(k)}$ can be obtained from those of iteration $k-1$ at cost comparable to one matrix-vector multiplication with A, $2k$ vector scalings and additions with vectors of length n and additional $\mathcal{O}(k^2)$ arithmetic operations.*

Proof. Computing the last column of $V_i^{(k)} R_i^{(k)}$ costs k vector scalings and additions with vectors of length n_i for $i = 1, 2$, which is comparable to k scalings and additions with vectors of length n. Multiplication of these last columns with the A_{ij} in (4.2) amounts to one matrix-vector multiplication with A. Orthogonalizing the two resulting blocks against all columns of $V_k^{(i)}$ costs again k scalings and additions of vectors of size n_1 and n_2, which corresponds to additional k such operations on vectors of length n. All other necessary updates as described before require $\mathcal{O}(k^2)$ operations. □

In the standard Arnoldi process, when $\eta^{(k)} = 0$, we know that we have reached the maximum size of the Krylov subspace, i.e., k is equal to the grade of the initial residual $r^{(0)}$, and that $A^{-1}b$ is contained in $\mathcal{K}^{(k)}(A, r^{(0)})$. Since by (4.3) we have $\eta_i^{(k)} = \rho_i^{(k)} \eta^{(k)}$, $i = 1, 2$, we see that the two-level orthogonal Arnoldi method also stops when $\eta^{(k)} = 0$. However, the reverse statement need not necessarily be true, i.e., we can have $\eta_i^{(k)} = 0$ for $i = 1, 2$ without having $\eta^{(k)} = 0$. This would represent a serious breakdown of the two-level orthogonal Arnoldi process. Of course, exact zeros rarely appear in a numerical computation, but near breakdowns should be dealt with appropriately. In our implementation, we simply chose to replace a block vector corresponding to some $\eta_i^{(k)} \approx 0$ by a vector with just random entries. This makes the book-keeping much easier, since then $d_i^{(k)} = k$ for all k and $i = 1, 2$, while keeping $V_\times^{(k)}$ as a subspace of our approximation space.

The full algorithm is summarized in Algorithm 1. We assume no deflation is necessary and no breakdown occurs for simplicity, but we can deal with this in practice in two ways. When $\widetilde{v}_i^{(k+1)}$ is (numerically) linear dependent, we can either set $v_i^{(k+1)}$ to some random vector and set $\eta_i^{(k)}$ to zero, or we can set $V_i^{(k+1)} = V_i^{(k)}$ and $\underline{H}_i^{(k)} = [H_i^{(k-1)} \;\; h_i^{(k)}]$. The former approach requires less bookkeeping, but the latter approach can save space and time. Another simplification compared to a practical implementation is the use of classical Gram–Schmidt for the orthogonalization, instead of repeated Gram–Schmidt or modified Gram–Schmidt. However, the algorithm does show how to avoid unnecessary recomputation of quantities. In particular, we avoid recomputing matrix-vector products by updating the products $W_{ij}^{(k)} = A_{ij} V_j^{(k)}$, $Z_{ij}^{(k,k)} = (V_i^{(k)})^* A_{ij} V_j^{(k)}$, and $Z_{ij}^{(k+1,k)} = (V_i^{(k+1)})^* A_{ij} V_j^{(k)}$. Since this updating approach requires more memory, it should only be used if that extra memory is available, and if matrix-vector products with A are sufficiently expensive.

Algorithm 1: Quadratic FOM (QFOM).

```
Input: $A_{11}$, $A_{12}$, $A_{21}$, $A_{22}$, $b_1$, $b_2$, $k_{\max}$, and $\tau$
$\underline{H}^{(0)} = [\,]$ and $\beta = (\|b_1\|^2 + \|b_2\|^2)^{-1/2}$
for $i = 1, 2$
    $\rho_i^{(1)} = \|b_i\|/\beta$ and $v_i^{(1)} = b_i/\rho_i^{(1)}$
    $\underline{H}_i^{(0)} = [\,]$, $R_i^{(1)} = [\rho_i^{(1)}]$, and $V_i^{(1)} = [v_i^{(1)}]$
for $k = 1$ to $k_{\max}$
    for $i = 1, 2$                                  /* Update matrix products.   */
        for $j = 1, 2$
            $w_{ij}^{(k)} = A_{ij} v_j^{(k)}$
            $W_{ij}^{(k)} = [W_{ij}^{(k-1)}\ w_{ij}^{(k)}]$
            $Z_{ij}^{(k,k)} = [Z_{ij}^{(k,k-1)}\ (V_i^{(k)})^* w_{ij}^{(k)}]$
    for $i = 1, 2$                                  /* Update $V_i^{(k+1)}$ and $\underline{H}_i^{(k)}$.   */
        $\widetilde{v}_i^{(k+1)} = W_{i1}^{(k)}(R_1^{(k)} e^{(k)}) + W_{i2}^{(k)}(R_2^{(k)} e^{(k)})$
        $h_i^{(k)} = (V_i^{(k)})^* \widetilde{v}_i^{(k+1)}$
        $\eta_i^{(k)} = \|\widetilde{v}_i^{(k+1)} - V_i^{(k)} h_h^{(k)}\|$
        $v_i^{(k+1)} = (\widetilde{v}_i^{(k+1)} - V_i^{(k)} h_h^{(k)})/\eta_i^{(k)}$
        $V_i^{(k+1)} = [V_i^{(k)}\ v_i^{(k+1)}]$ and $\underline{H}_i^{(k)} = \begin{bmatrix} H_i^{(k-1)} & h_i^{(k)} \\ 0^T & \eta_i^{(k)} \end{bmatrix}$
        for $j = 1, 2$
            $Z_{ij}^{(k+1,k)} = [Z_{ij}^{(k,k)};\ (v_i^{(k+1)})^* W_{ij}^{(k)}]$
    /* Update $\underline{H}^{(k)}$ and $R_i^{(k+1)}$.                                  */
    $h^{(k)} = (R_1^{(k)})^* h_1^{(k)} + (R_2^{(k)})^* h_2^{(k)}$
    for $i = 1, 2$
        $\widetilde{r}_i^{(k+1)} = h_i^{(k)} - R_i^{(k)} h^{(k)}$
    $\eta^{(k)} = (\|\widetilde{r}_i^{(k+1)}\|^2 + |\eta_1^{(k)}|^2 + \|\widetilde{r}_i^{(k+1)}\|^2 + |\eta_2^{(k)}|^2)^{1/2}$
    for $i = 1, 2$
        $r_i^{(k+1)} = \widetilde{r}_i^{(k+1)}/\eta^{(k)}$, $\rho_i^{(k+1)} = \eta_i^{(k)}/\eta^{(k)}$, and $R_i^{(k+1)} = \begin{bmatrix} R_i^{(k)} & r_i^{(k+1)} \\ 0 & \rho_i^{(k+1)} \end{bmatrix}$
    /* Compute the approximation $x_{\mathtt{qfom}}^{(k)}$ and the residual
        $r_{\mathtt{qfom}}^{(k)}$.                                                  */
    $H_\times^{(k)} = \begin{bmatrix} Z_{11}^{(k,k)} & Z_{12}^{(k,k)} \\ Z_{21}^{(k,k)} & Z_{22}^{(k,k)} \end{bmatrix}$ and $b_\times^{(k)} = \beta \begin{bmatrix} R_1^{(k)} e^{(k)} \\ R_2^{(k)} e^{(k)} \end{bmatrix}$
    $\begin{bmatrix} c_{\mathtt{qfom}}^{(k)} \\ d_{\mathtt{qfom}}^{(k)} \end{bmatrix} = (H_\times^{(k)})^{-1} b_\times^{(k)}$ and $x_{\mathtt{qfom}}^{(k)} = \begin{bmatrix} V_1^{(k)} c_{\mathtt{qfom}}^{(k)} \\ V_2^{(k)} d_{\mathtt{qfom}}^{(k)} \end{bmatrix}$
    $r_{\mathtt{qfom}}^{(k)} = \begin{bmatrix} b_1 - W_{11}^{(k)} c_{\mathtt{qfom}}^{(k)} - W_{12}^{(k)} d_{\mathtt{qfom}}^{(k)} \\ b_2 - W_{21}^{(k)} c_{\mathtt{qfom}}^{(k)} - W_{22}^{(k)} d_{\mathtt{qfom}}^{(k)} \end{bmatrix}$
    if $\|r_{\mathtt{qfom}}^{(k)}\| \le \tau\beta$ then
        return $x_{\mathtt{qfom}}^{(k)}$
return $x_{\mathtt{qfom}}^{(k_{\max})}$
```

From the pseudocode of the algorithm we can determine the computational cost per iteration as follows. We count one matrix-vector multiplication with each of the blocks A_{11}, A_{12}, A_{21}, and A_{22}, which equals one matrix-vector multiplication with A. Then we have an orthogonalization cost of $\mathcal{O}((n_1+n_2)k) = \mathcal{O}(nk)$, which equals the orthogonalization cost in the standard Arnoldi process. Updating the Z_{ij} costs $\mathcal{O}(nk)$ floating-point operations per iteration, but does not have an equivalent cost in Arnoldi. The same is true for updating the matrices $\underline{H}^{(k)}$ and $R_i^{(k+1)}$ for $i = 1, 2$, although the cost is limited to $\mathcal{O}(k)$ flops in this case. Computing $c^{(k)}_{\mathtt{qfom}}$ and $d^{(k)}_{\mathtt{qfom}}$ takes $\mathcal{O}(k^3)$ floating-point operations, while computing the approximation $x^{(k)}_{\mathtt{qfom}}$ and its residual $r^{(k)}_{\mathtt{qfom}}$ require $\mathcal{O}(nk)$. Clearly, computing the approximation and its residual is expensive, but there is no need to do it in every iteration. For example, in a restarted version of the QFOM algorithm, we may decide to compute them only once per restart, after the inner loop reaches $k_{\max}$. When we add everything together, we see that QFOM has the same asymptotic cost as FOM, although QFOM does require more memory.

With minor changes, we can change the code of Algorithm 1 to compute the QQGMRES approximation instead of the QFOM approximation. One downside of QQGMRES is that we cannot guarantee that its approximation, or even the residual norm of its approximation, is better than that of GMRES. We can remedy this problem by interpolating between the GMRES and the QQGMRES solution. Let $r^{(k)}_{\mathtt{gmres}} = b - Ax^{(k)}_{\mathtt{gmres}}$ and $r^{(k)}_{\mathtt{qqgmr}} = b - Ax^{(k)}_{\mathtt{qqgmr}}$, then

$$\begin{aligned}\|b - A(\alpha x^{(k)}_{\mathtt{gmres}} + (1-\alpha)x^{(k)}_{\mathtt{qqgmr}})\|^2 &= \|\alpha r^{(k)}_{\mathtt{gmres}} + (1-\alpha) r^{(k)}_{\mathtt{qqgmr}}\|^2 \\ &= \alpha^2\|r^{(k)}_{\mathtt{gmres}} - r^{(k)}_{\mathtt{qqgmr}}\|^2 + 2\alpha(\Re\{(r^{(k)}_{\mathtt{gmres}})^* r^{(k)}_{\mathtt{qqgmr}}\} - \|r^{(k)}_{\mathtt{qqgmr}}\|^2) + \|r^{(k)}_{\mathtt{qqgmr}}\|^2.\end{aligned}$$

Hence, the residual norm of the interpolated approximation is minimized for

$$\alpha_{\mathtt{opt}} = \frac{\|r^{(k)}_{\mathtt{qqgmr}}\|^2 - \Re\{(r^{(k)}_{\mathtt{gmres}})^* r^{(k)}_{\mathtt{qqgmr}}\}}{\|r^{(k)}_{\mathtt{gmres}} - r^{(k)}_{\mathtt{qqgmr}}\|^2}$$

if $r^{(k)}_{\mathtt{gmres}} \neq r^{(k)}_{\mathtt{qqgmr}}$. The residual norm of the approximation $x^{(k)}_{\mathtt{opt}}$ corresponding to $\alpha_{\mathtt{opt}}$ is

$$\|r_{\mathtt{opt}}\|^2 = \frac{\|r^{(k)}_{\mathtt{gmres}}\|^2 \|r^{(k)}_{\mathtt{qqgmr}}\|^2 - \Re\{(r^{(k)}_{\mathtt{gmres}})^* r^{(k)}_{\mathtt{qqgmr}}\}^2}{\|r^{(k)}_{\mathtt{gmres}} - r^{(k)}_{\mathtt{qqgmr}}\|^2},$$

and satisfies $\|r_{\mathtt{opt}}\| \leq \min\{\|r_{\mathtt{gmres}}\|, \|r_{\mathtt{qqgmr}}\|\}$. Obtaining $x^{(k)}_{\mathtt{opt}}$ requires the computation of the GMRES iterate $x^{(k)}_{\mathbf{gmres}}$. This can be done without additional matrix-vector multiplications, since by (3.2) we have

$$\underline{H}^{(k)} := (V^{(k+1)})^* A V^{(k)} = \begin{bmatrix} (R_1^{(k+1)})^* & (R_2^{(k+1)})^* \end{bmatrix} \underline{H}_\times^{(k)} \begin{bmatrix} R_1^{(k)} \\ R_2^{(k)} \end{bmatrix}.$$

So $\underline{H}^{(k)}$ can be retrieved with cost $\mathcal{O}(k^3)$. If we compute the GMRES coefficients $\xi_k \in \mathbb{C}^k$ in the usual way from the least squares problem with $\underline{H}^{(k)}$ and obtain $x^{(k)}_{\mathbf{gmres}}$ as

$$x^{(k)}_{\mathbf{gmres}} = x^{(0)} + V^{(k)}\xi_k = x^{(0)} + \begin{bmatrix} V_1^{(k)} & 0 \\ 0 & V_2^{(k)} \end{bmatrix} \left(\begin{bmatrix} (R_1^{(k)}) \\ R_2^{(k)} \end{bmatrix} \xi_k \right)$$

we see that the additional cost for getting $x^{(k)}_{\mathbf{gmres}}$ and thus $x^{(k)}_{\mathtt{opt}}$ is comparable to that for obtaining $x^{(k)}_{\mathbf{qqgmr}}$ and of order $\mathcal{O}(nk) + \mathcal{O}(k^3)$.

5. Numerical experiments.

5.1. The Hain-Lüst operator. Hain-Lüst operators appear in magnetohydrodynamics [6]. Their spectral properties, in particular their quadratic numerical range, were investigated in a series of papers, e.g., in [11, 15, 16]. We consider the Hain-Lüst operator

$$\mathcal{A} = \begin{bmatrix} -\mathcal{L} & I \\ I & q \end{bmatrix}$$

defined on $L^2([0,1]) \times L^2([0,1])$ where $\mathcal{L} = d^2/dx^2$ is the Laplace operator on $[0,1]$ with zero Dirichlet boundary conditions, I is the identity operator, and q denotes multiplication by the function $q(x) = -3 + 2e^{2\pi i x}$. The domain of $\mathcal{A}$ is

$$D(\mathcal{A}) = (H^2([0,1]) \cap H_0^1([0,1])) \times L^2([0,1]).$$

We consider a discretization of $\mathcal{A}$, approximating function values at an equispaced grid for both blocks, i.e., we take $x_j = jh$, $j = 0, \dots, N+1$, $h = 1/(N+1)$ and obtain, using finite differences, the discretized Hain-Lüst operator

$$A = \begin{bmatrix} \frac{1}{h^2} L & I \\ I & Q \end{bmatrix} \in \mathbb{C}^{2N \times 2N},$$

with $L = \text{tridiag}(-1, 2, -1) \in \mathbb{C}^{N \times N}$ and $Q = -3I + 2\,\text{diag}(e^{2h\pi i}, \dots, e^{2hN\pi i}) \in \mathbb{C}^{N \times N}$; see [16] for more details.

Note that $\frac{1}{h^2} L$ is Hermitian and that Q is normal, so the numerical ranges of these diagonal blocks of A satisfy

$$\begin{aligned} W_1 := W(\tfrac{1}{h^2} L) \; &= \; \tfrac{1}{h^2}[2 - 2\cos(\pi h), 2 + 2\cos(\pi h)] \; =: \; [\alpha_{\min}(h), \alpha_{\max}(h)], \\ W_2 := W(Q) \; &= \; \text{conv}\{-3 + 2e^{2\pi h j}, j = 1, \dots, N\} \subseteq C(-3, 2), \end{aligned}$$

where $C(-3, 2)$ is the circle with center -3 and radius 2. Since both numerical ranges W_1 and W_2 are contained in the convex set $W(A)$ we see that $0 \in W(A)$. The following argumentation shows that, with the possible exception of very large values for h, we have $0 \notin W^2(A)$: any $\lambda \in W^2(A)$ satisfies

$$(\lambda - x_1^* \tfrac{1}{h^2} L x_1)(\lambda - x_2^* Q x_2) = (x_1^* x_2)(x_2^* x_1), \tag{5.1}$$

for some x_1, x_2 with $\|x_1\| = \|x_2\| = 1$. Assume that λ lies within the strip $a < \Re(\lambda) < b$ with $-1 < a < 0$ and $0 < b < \alpha_{\min}(h)$. Then we have $d(\lambda, W_1) > \alpha_{\min}(h) - b$ as well as $d(\lambda, W_2) > a + 1$ for the distances of λ to the sets W_1, W_2. Taking absolute values in (5.1) and using the bound $|x_1^* x_2| \leq 1$ we thus see that the λ from this strip cannot be in $W^2(A)$ if $(a+1)(\alpha_{\min}(h) - b) > 1$. This is the case, for example, if $b < \alpha_{\min}(h) - 2$ and $a > -\frac{1}{2}$. Note that $\lim_{h \to 0} \alpha_{\min}(h) = \pi^2$.

In all our examples we chose the right hand side b such that e, the vector of all ones, is the solution, imposing boundary conditions also equal to 1 for the Dirichlet part. Our initial guess is always $x^0 = 0$. Figure 5.1 shows convergence plots for FOM, GMRES, QFOM, QQGMRES and the interpolated QQGMRES method as described at the end of Section 4. The figure displays the relative norm of the residual as a function of the invested matrix-vector multiplications. In the left part we took $N = 1\,023$, the right part is for $N = 16\,383$. We restarted every method after $m = 50$ iterations to avoid that the arithmetic work and the storage related with the (two-level) Arnoldi process becomes too expensive. Note that the figure displays the residual norms at the end of each cycle only, which makes the convergence of some of the

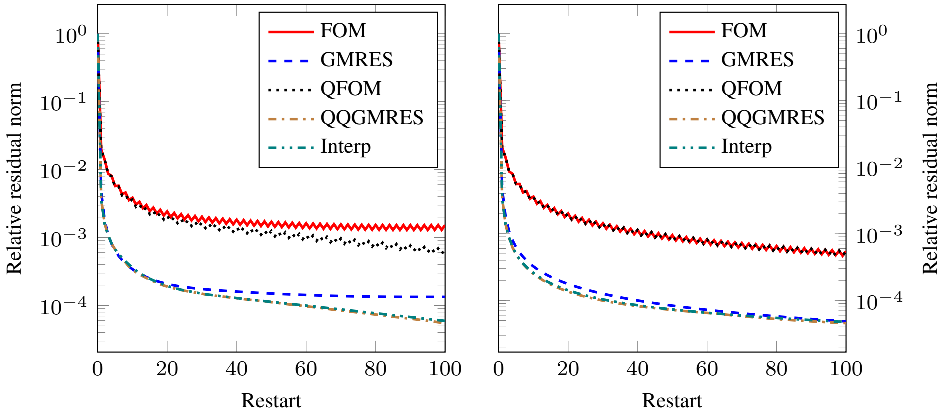

FIG. 5.1. *Convergence plots for the discretized Hain-Lüst operator: $N = 1\,023$ (left) and $N = 16\,383$ (right).*

methods, in particular FOM, to appear smoother than it actually is. Two major observations can be made: On the one side, the FOM type methods yield significantly larger residuals than the GMRES type methods. For $N = 1\,023$, the "quadratic methods" still make progress in the later cycles while their "non-quadratic" counter parts then basically stagnate. There is no such difference visible for dimension $N = 16\,383$; convergence for all methods is very slow.

In a second numerical experiment we therefore report results of a geometric multigrid method as an attempt to cope with large condition numbers. For a given discretization with step size $h = 1/(N+1)$ with $N+1 = 2^k$ we construct the system at the next coarser level to be the discretizaton with $h_c = 2h = 1/(N_c + 1)$ with $N_c + 1 = 2^{k-1}$. We stop descending the grid hierarchy when we reach $N = 7$, where we solve the corresponding 14×14 system by explicit inversion of A. Interpolation between two levels of the grid hierarchy is done using standard linear interpolation from the neighboring grid points; restriction is the standard adjoint of interpolation. For the smoothing iteration we test one or five steps of standard GMRES versus one or two steps of QFOM. We always performed V-cycles with pre-smoothing. In addition, we report results using the Kaczmarz smoother [24], the standard smoother for indefinite and non-symmetric problems. The left part of Figure 5.2 gives the resulting convergence plots for the multigrid methods for $N = 1\,023$, the right part for $N = 16\,383$.

From these plots it is apparent that QFOM is a well-working smoothing iteration for the multigrid method, whereas GMRES is not, even not for larger numbers of smoothing steps per iteration. Using one step of Kaczmarz smoothing, we obtain a slow, but convergent method. Since Kaczmarz works on the normal equations, one smoothing step is roughly twice as costly as one step of the other smoothing methods considered here. If we want to obtain a convergence speed similar to that with one step of QFOM smoothing, our numerical experiments showed that we have to invest more than ten Kaczmarz smoothing iterations for both dimensions. The bottom line is that for both dimensions considered here, using one step of QFOM smoothing is by far the best method, with GMRES smoothing resulting in stagnation and Kaczmarz smoothing being at least one order of magnitude more costly. As a complement to these findings, Figure 5.3 illustrates the mesh size independence of the convergence behavior of the multigrid method with QFOM smoothing. It shows that the number of iterations required to reduce the initial residual by a factor of 10^{-12} is basically independent of h.

5.2. The Schwinger model. Our second example is the Schwinger model in two-dimensions that arises in computations of quantum electrodynamics (QED). QED models the

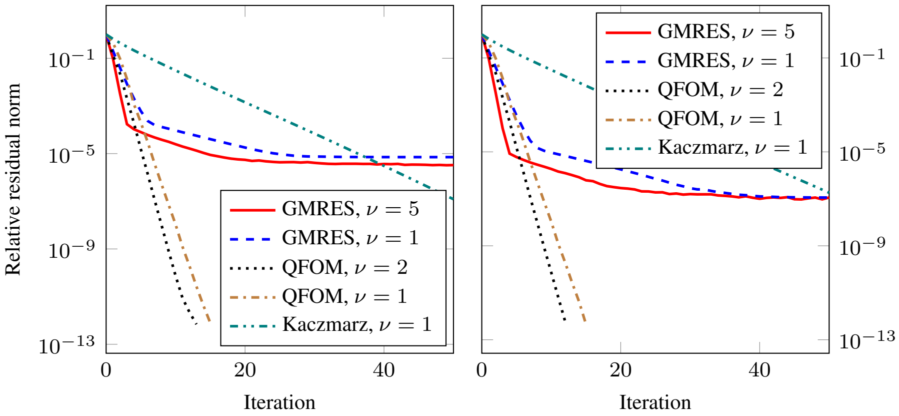

FIG. 5.2. *Convergence plots for geometric multigrid for the Hain-Lüst operator for QFOM, GMRES, and Kaczmarz smoothing and different numbers of smoothing steps* ν; $N = 1\,023$ *(left),* $N = 16\,383$ *(right).*

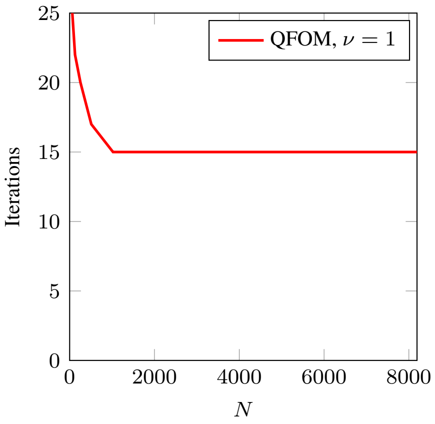

FIG. 5.3. *Number of multigrid iterations needed to reduce the initial residual by* 10^{-12} *as a function of* N.

interactions of electrons and photons and is oftentimes used as a simpler model problem for the 4-dimensional problems of quantum chromodynamics (QCD). It is a quantum field theory, meaning that physical quantities arise as expected values of solutions of partial differential equations whose coefficients are coming from the quantum background field, i.e., they are stochastic quantities obeying a given distribution. The Schwinger model is a discretization of the Dirac equation

$$\mathcal{D}\psi = (\sigma_1 \otimes (\partial_x + A_x) + \sigma_2 \otimes (\partial_y + A_y))\,\psi = \varphi,$$

on a regular, 2-dimensional $N \times N$ Cartesian lattice, where the spin structure[*] is encoded by the Pauli matrices

$$\sigma_1 = \begin{bmatrix} & 1 \\ 1 & \end{bmatrix}, \quad \sigma_2 = \begin{bmatrix} & i \\ -i & \end{bmatrix} \quad \text{and} \quad \sigma_3 = \begin{bmatrix} 1 & \\ & -1 \end{bmatrix}$$

[*] The σ-matrices are generators of a Clifford algebra and arise in the derivation of the Dirac equation from the Klein-Gordon equation. They give rise to the internal spin (i.e., angular momentum) degrees of freedom of the fields ψ [4]. Note that although our discussion is limited to this particular choice of generators, all the results that follow extend to any other of the admissible choices of the σ-matrices.

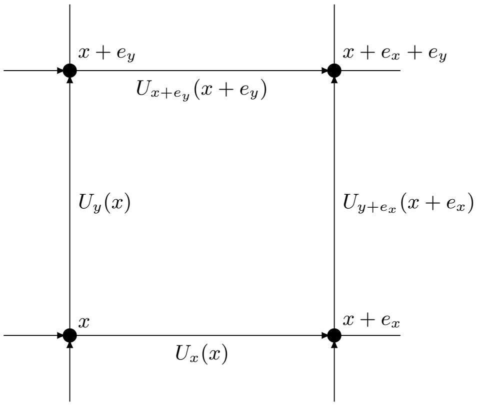

FIG. 5.4. *Naming conventions in the Schwinger model.*

and A_μ encodes the background gauge field. In the Schwinger model we have $A_\mu \in \mathbb{R}$. Using a central covariant finite difference discretization for the first order derivatives, and introducing a scaled second-order stabilization term one writes the action of the discretized operator $D \in \mathbb{C}^{2N^2 \times 2N^2}$ of the Schwinger model at any lattice site x on a spinor $\psi(x) \in \mathbb{C}^2$ as

$$\left\{ \begin{aligned} (D\psi)(x) \;&=\; (m_0+2)\,\psi(x) \\ &\quad + \frac{1}{2} \sum_{\mu \in \{x,y\}} \left((I-\sigma_\mu) \otimes U_\mu(x)\right) \psi(x+e_\mu) \\ &\quad + \frac{1}{2} \sum_{\mu \in \{x,y\}} \Big((I+\sigma_\mu) \otimes \overline{U_\mu(x-e_\mu)}\Big)\, \psi(x-e_\mu). \end{aligned} \right. \tag{5.2}$$

In here U_μ corresponds to a discrete version of the stochastically varying gauge field with $U_\mu(x) \in \mathbb{C}$, $|U_\mu(x)| = 1$ for all x, and m_0 sets the mass of the simulated theory. The naming convention of this formula is depicted in Figure 5.4, and we refer to the textbook [4], e.g., for further details. The canonical 2×2 block structure of the Schwinger model matrix arises from the spin structure: we reorder the unknowns in ψ according to spin, i.e., we take

$$\psi = \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix},$$

where $\psi_1 \in \mathbb{C}^{N^2}$ collects all the spin 1 components $\psi_1(x)$ of $\psi(x) = \left[\begin{smallmatrix} \psi_1(x) \\ \psi_2(x) \end{smallmatrix}\right] \in \mathbb{C}^2$ at all lattice sites, and similarly for ψ_2. Then the reordered discretized Schwinger model matrix, acting on the reordered vector $\left[\begin{smallmatrix} \psi_1(x) \\ \psi_2(x) \end{smallmatrix}\right]$, is given as

$$D = \begin{bmatrix} A & B \\ -B^* & A \end{bmatrix}.$$

Here, the diagonal blocks A correspond to the discretized second order stabilization term and are thus called gauge Laplace operators, while the off-diagonal blocks B correspond to the central finite covariant difference discretization of the Dirac equation. Using (5.2) we see that the action of the blocks A and B on a vector ψ_1, ψ_2 is given as

$$\begin{aligned}(A\psi_1)(x) = (m_0+2)\psi_1(x) &- \frac{1}{2}\sum_{\mu\in\{x,y\}} U_\mu(x)\psi_1(x+e_\mu)\\ &- \frac{1}{2}\sum_{\mu\in\{x,y\}} \overline{U_\mu(x-e_\mu)}\psi_1(x-e_\mu),\\ (B\psi_2)(x) = \; &-\frac{1}{2}\left(U_x(x)\psi_1(x+e_x) + i\cdot U_y(x)\psi_1(x+e_y)\right)\\ &+\frac{1}{2}\left(\overline{U_x(x-e_x)}\psi_1(x-e_x) - i\cdot\overline{U_y(x-e_y)}\psi_1(x-e_y)\right).\end{aligned}$$

From this we see that the mass parameter m_0 induces a shift by a multiple of the identity in A, which we make explicit in writing $A = A_0 + m_0 I$.

In our tests we consider the "symmetrized" operator $Q := \Sigma_3 D$ with $\Sigma_3 = \sigma_3 \otimes I_{N\cdot N}$. Due to $A^* = A, B^* = -B$ this operator

$$Q = \begin{bmatrix} A & B \\ B^* & -A \end{bmatrix} = \begin{bmatrix} A_0 + m_0 I & B \\ B^* & -A_0 - m_0 I \end{bmatrix}$$

is Hermitian, but indefinite.

The quadratic range $W^2(Q)$ has two connected components to the left and right of 0 on the real axis, provided $m_0 > -\alpha_{\min}$, the smallest eigenvalue of A_0. This can be seen as follows: Let $x_1, x_2 \in \mathbb{C}^{N\times N}$ be two normalized vectors and let

$$\begin{bmatrix} x_1^* A x_1 & x_1^* B x_2 \\ x_2^* B^* x_1 & -x_2^* A x_2 \end{bmatrix} =: \begin{bmatrix} \alpha_1 & \beta \\ \overline{\beta} & -\alpha_2 \end{bmatrix}.$$

Then any eigenvalue λ of this matrix satisfies

$$\begin{aligned} (\lambda-\alpha_1)(\lambda+\alpha_2) &= |\beta|^2 \\ \Longrightarrow (\Re(\lambda)-\alpha_1)(\Re(\lambda)+\alpha_2) &= |\beta|^2 + \Im(\lambda)^2. \end{aligned}$$

The last equality cannot be satisfied if $-\alpha_2 < \Re(\lambda) < \alpha_1$. In particular, if $m_0 > -\alpha_{\min}$, the equality cannot be satisfied if $|\Re(\lambda)| < m_0 + \alpha_{\min}$, since $\alpha_1, \alpha_2 \geq m_0 + \alpha_{\min}$.

For our tests we use a gauge configuration obtained by a heat bath algorithm excluding the fermionic action, which results in the smallest eigenvalue $\alpha_{\min}$ of A_0 being approximately 0.11. Figure 5.5 reports results for two different choices of m_0. As in the first example we perform a restart after every 50 iterations. The first choice for m_0 is $m_0 = -0.1 > -\alpha_{\min}$, so that the quadratic range indeed has two connected components with a gap around 0. The second is $m_0 = -0.21 < -\alpha_{\min}$, so that $W^2(Q)$ consists of only one component containing 0. The figure shows that a marked improvement can be observed for the "quadratic" methods if the quadratic range consists indeed of two different connected components (left plot), whereas this advantage is lost to a large extent for the second choice for m_0, where $W^2(Q)$ does not indicate a spectral gap (right plot). In this case, the system is also severely ill-conditioned, so that the convergence of all methods considered is much slower. As a reference, we also included convergence plots for the MINRES method [19], an algorithmic variant of GMRES tailored towards Hermitian and indefinite systems, which relies on short recurrences and thus does not require restarts. Very remarkably, for $m_0 = -0.1$ even though we performed about 40 restarts, QFOM performs very similarly to MINRES, whereas MINRES outperforms all other methods for $\alpha = -0.21$. We also note that for this example and for both choices for

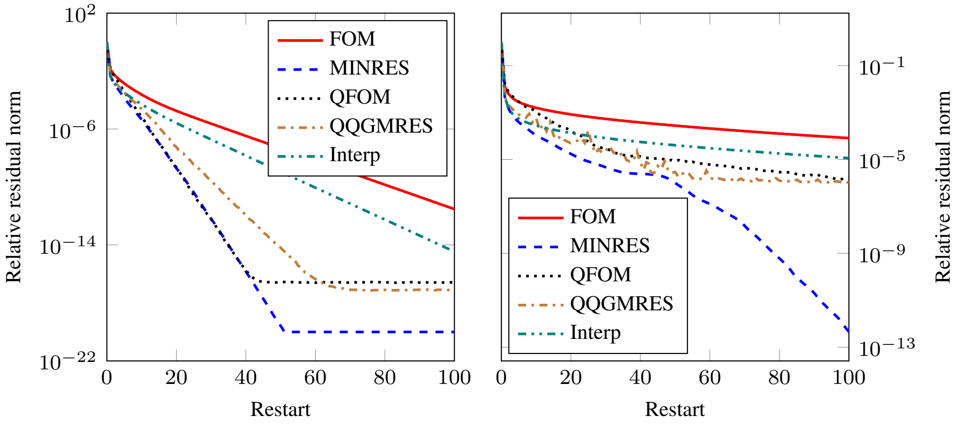

FIG. 5.5. *Convergence plots for the Schwinger model,* $\alpha_{\min} \approx 0.11$*,* $N = 128$*. Left:* $m_0 = -0.1 > -\alpha_{\min}$*. Right:* $m_0 = -0.21 < -\alpha_{\min}$*.*

m_0, interpolated QQGMRES does not differ substantially from standard (restarted) GMRES. Without showing the corresponding convergence plots, let us at least mention that when decreasing m_0 from -0.1 to -0.21 we observe for a long time a convergence behavior very similar to that for the largest value -0.1, even when m_0 is already smaller than $-\alpha_{\min}$.

Note that the symmetrized Schwinger model matrix $Q = \left[\begin{smallmatrix} C & B \\ B^* & -C \end{smallmatrix}\right]$ is an example of a saddle point problem, and we showed that $W^2(Q) \cap (-\alpha, \alpha) = \emptyset$ if C is Hermitian and positive definite with α its smallest eigenvalue. This can be extended to non-Hermitian matrices of the form $T = \left[\begin{smallmatrix} A & B \\ B^* & -C \end{smallmatrix}\right]$, where A and C are such that $\Re(W(A)) \geq \alpha > 0$ and $\Re(W(C)) \geq \gamma > 0$ by showing that $W^2(T)$ has empty intersection with the strip $\{z \in \mathbb{C} : -\gamma < \Re(z) < \alpha\}$. So QFOM and QQGMRES are adapted to such kind of saddle point problems. On the other hand, if $T = \left[\begin{smallmatrix} A & B \\ B^* & 0 \end{smallmatrix}\right]$, it is easy to show that $0 \in W(T)$ for any choice of A, so that the basic property $0 \notin W^2(T)$ motivating QFOM and QQGMRES is not met.

Acknowledgement. We would like to thank the two anonymous referees for valuable comments and suggestions.

REFERENCES

[1] Z. Bai and Y. Su, *SOAR: a second-order Arnoldi method for the solution of the quadratic eigenvalue problem*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 640–659.

[2] M. Eiermann and O. G. Ernst, *Geometric aspects of the theory of Krylov subspace methods*, Acta Numer., 10 (2001), pp. 251–312.

[3] S. C. Eisenstat, H. C. Elman, and M. H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 20 (1983), pp. 345–357.

[4] C. Gattringer and C. B. Lang, *Quantum Chromodynamics on the Lattice: An Introductory Presentation*, Springer, Berlin, 2010.

[5] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.

[6] K. Hain and R. Lüst, *Zur Stabilität zylindersymmetrischer Plasmakonfigurationen mit Volumenströmen*, Z. Naturf., 13a (1958), pp. 936–940.

[7] N. J. Higham, *Functions of Matrices*, SIAM, Philadelphia, 2008.

[8] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed., Cambridge University Press, Cambridge, 2013.

[9] D. Kressner and J. E. Roman, *Memory-efficient Arnoldi algorithms for linearizations of matrix polynomials in Chebyshev basis*, Numer. Linear Algebra Appl., 21 (2014), pp. 569–588.

[10] H. Langer, A. Markus, V. Matsaev, and C. Tretter, *A new concept for block operator matrices: the quadratic numerical range*, Linear Algebra Appl., 330 (2001), pp. 89–112.

[11] H. LANGER AND C. TRETTER, *Spectral decomposition of some nonselfadjoint block operator matrices*, J. Operator Theory, 39 (1998), pp. 339–359.

[12] J. LIESEN AND Z. STRAKOŠ, *Krylov Subspace Methods: Principles and Analysis*, Oxford University Press, Oxford, 2013.

[13] D. LU, Y. SU, AND Z. BAI, *Stability analysis of the two-level orthogonal Arnoldi procedure*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 195–214.

[14] K. MEERBERGEN AND J. PÉREZ, *Mixed forward-backward stability of the two-level orthogonal Arnoldi method for quadratic problems*, Linear Algebra Appl., 553 (2018), pp. 1–15.

[15] A. MUHAMMAD AND M. MARLETTA, *Approximation of the quadratic numerical range of block operator matrices*, Integral Equations Operator Theory, 74 (2012), pp. 151–162.

[16] ———, *A numerical investigation of the quadratic numerical range of Hain-Lüst operators*, Int. J. Comput. Math., 90 (2013), pp. 2431–2451.

[17] L. PENG AND K. MOHSENI, *Symplectic model reduction of Hamiltonian systems*, SIAM J. Sci. Comput., 38 (2016), pp. A1–A27.

[18] Y. SAAD, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.

[19] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.

[20] 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.

[21] G. STARKE, *Field-of-values analysis of preconditioned iterative methods for nonsymmetric elliptic problems*, Numer. Math., 78 (1997), pp. 103–117.

[22] L. N. TREFETHEN AND M. EMBREE, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, Princeton, 2005.

[23] C. TRETTER, *Spectral Theory of Block Operator Matrices and Applications*, Imperial College Press, London, 2008.

[24] U. TROTTENBERG, C. OOSTERLEE, AND A. SCHÜLLER, *Multigrid*, Academic Press, New York, 2001.