Eötvös Loránd University
Faculty of Science

Comparison of iterative methods for discretized nonsymmetric elliptic problems

Math Project III.

Lados Bálint István

Applied Mathematics MSc
Applied Analysis specialization

Supervisor:

Karátson János

Department of Applied Analysis and
Computational Mathematics

ELTE
EÖTVÖS LORÁND
TUDOMÁNYEGYETEM

Budapest, 2024

1 Introduction

In fluid dynamics, fundamental models of the stationary flow of incompressible fluids are given by PDE systems such as Navier–Stokes equations. Stationary convection-diffusion equations can be considered as the elementary building blocks of the linearized version of such models. These are in general nonsymmetric elliptic partial differential equations, endowed with boundary conditions, which describe convective and diffusive effects simultaneously. A practically relevant subclass of these is formed by the so-called convection-dominated problems, where the impact of diffusion is less significant compared to the convection of the modeled fluid.

The analytical solution of such problems cannot be calculated in general, so instead we use numerical methods to approximate the solution to the desired accuracy. This is usually achieved with the finite difference (FDM) or the finite element method (FEM). The discretization process results in a system of linear equations that can be solved by a nonsymmetric iterative method, such as the Conjugate Gradient method applied to the normal equation (CGN) or the Generalized Conjugate Residual method (GCR).

It has been shown in [1] by earlier authors that there is no single "best" nonsymmetric iterative method that we know of, that is, which would dominate all the others with respect to the number of iterative steps required to decrease the error of the approximation below a given threshold. This work aims to study this question in the practically important special case when the matrices come from discretized nonsymmetric elliptic problems.

In the preliminaries, we formulate the studied nonsymmetric elliptic problems, review the methods used for generating the discretization matrix, and present the mentioned nonsymmetric iterative methods along with some well-known linear and superlinear convergence estimates.

After this section, the new results are presented. In this thesis there are both theoretical and numerical achievements. Detailed numerical experiments are conducted in MATLAB for various nonsymmetric elliptic test problems with FDM, FEM and SDFEM discretizations to see which iterative method prevails under which conditions. These experiments suggest that for our special class of problems, depending on the coefficients of the PDE, we can still forecast (or at least explain) which method will converge faster. In order to study the rate of convergence, the linear and superlinear convergence estimates are compared theoretically and numerically.

The theoretical results in the linear case include the comparison of the estimates using the root of a cubic function, and its application for giving a computable upper bound in case of FEM discretization, and a lower and upper bound in case of SDFEM discretization for the parameter of the convection term, in which case the linear estimation of CGN is better than that of GCR. The superlinear estimations are compared for a constant vector field in the limiting case when the diffusion parameter is zero, and an upper bound is obtained in case of SDFEM discretization.

2 Nonsymmetric elliptic problems and their numerical solution

Let us consider the following boundary value problem on the domain $\Omega = (0,1) \times (0,1)$ with functions $f : \Omega \to \mathbb{R}$ and $\mathbf{w} : \overline{\Omega} \to \mathbb{R}^2$:

$$\begin{cases} -\alpha \Delta u + \mathbf{w} \cdot \nabla u = f \\ u|_{\partial\Omega} = 0 \end{cases} \tag{1}$$

In this convection-diffusion problem, the first-order nonsymmetric term $\mathbf{w} \cdot \nabla u$ describes convection, while the second-order term $-\alpha \Delta u$ determines the diffusion process.

Suppose that the following assumptions are satisfied:

(i) $\alpha > 0$ is a constant;

(ii) $\mathbf{w} \in C^1(\overline{\Omega}, \mathbb{R}^2)$ and $\operatorname{div}(\mathbf{w}) = 0$;

(iii) $f \in L^2(\Omega)$.

These conditions ensure (see [2], sec. 8.1) that the given boundary value problem has a unique weak solution, which means that $\exists!\ u \in H_0^1(\Omega)$ such that

$$\int_\Omega (\alpha \nabla u \cdot \nabla v + (\mathbf{w} \cdot \nabla u) v) = \int_\Omega f v \qquad (\forall v \in H_0^1(\Omega)). \tag{2}$$

2.1 Discretization methods

In this project, three dicretization methods are considered: FDM with the second-order central scheme, standard FEM with first-order Courant elements, and the SDFEM, that is a stabilized version of the standard FEM. We perform one of the discretization methods on the problem in (1) to give a finite dimensional approximation. The SDFEM is used especially when the problem is convection-dominated, i.e. α is a small positive constant (e.g. 10^{-2}), in which case the standard FEM converges slowly. The idea as described in [3] is that we extend the weak form with a stabilizing term containing a parameter $\delta > 0$ for which the usual choice is $\delta = O(h)$. If we chose $\delta := 0$, we would get back the standard FEM. The discretization results in a system of linear equations $Au = b$. The components of matrix A are $a_{ij} := \int_\Omega \left(\alpha \nabla \phi_j \cdot \nabla \phi_i + \left(\mathbf{w} \cdot \nabla \phi_j\right) \phi_i + \delta (\mathbf{w} \cdot \nabla \phi_j)(\mathbf{w} \cdot \nabla \phi_i)\right)$, and vector b has the components $b_i := \int_\Omega f\,(\phi_i + \delta \mathbf{w} \cdot \nabla \phi_i)$. The three discretization methods (FDM, standard FEM and SDFEM) have been implemented in MATLAB ([9]).

2.2 Nonsymmetric iterative methods

In order to determine the numerical solution, we have to solve the system of linear equations $Au = b$ obtained from the chosen discretization. Matrix A is not symmetric in general, so we need to use nonsymmetric iterative methods. In this thesis, we compare two widely used iterative methods: the Conjugate Gradient method applied to the normal equation (CGN) and the Generalized Conjugate Residual method (GCR). The basic algorithms can be found e.g. in [5].

These algorithms would be slow if we directly applied them to the original system of equations, so we rather solve the preconditioned system $S^{-1}Au = S^{-1}b$, where the preconditioner matrix S will be the symmetric part of A, i.e. $S := \frac{A+A^T}{2}$. As pointed out in [3], the components of S are

$$s_{ij} = \langle \phi_j, \phi_i \rangle_S := \int_\Omega \left(\alpha \nabla \phi_j \cdot \nabla \phi_i + \delta (\mathbf{w} \cdot \nabla \phi_j)(\mathbf{w} \cdot \nabla \phi_i)\right).$$

Matrix S is symmetric positive definite, so we can consider the energy inner product $\langle x, y \rangle_S := \langle Sx, y \rangle$ and the corresponding S-norm $\|x\|_S := \sqrt{\langle x, x \rangle_S}$, which will appear in the algorithms and the convergence estimates as well.

2.3 Linear and superlinear convergence estimates

The S-norm of the residual error vector $r_k = S^{-1}Au_k - S^{-1}b$ in the kth iterative step is bounded linearly (see [3]), which depends on the coercivity bound m and the S-norm of $S^{-1}A$:

$$\begin{aligned} m &:= \inf\left\{\left\langle S^{-1}Ac, c\right\rangle_S : \ \|c\|_S = 1\right\} = \inf\{\langle Ac, c\rangle : \ \|c\|_S = 1\} > 0 \\ M &:= \|S^{-1}A\|_S = \sup\{\langle Ac, d\rangle : \ \|c\|_S = \|d\|_S = 1\} \end{aligned} \tag{3}$$

The linear bound for the residual norm regarding the CGN and GCR methods is as follows:

$$\textit{Linear estimation of CGN:} \qquad \left(\frac{\|r_k\|_S}{\|r_0\|_S}\right)^{\frac{1}{k}} \leq 2^{\frac{1}{k}} \frac{M-m}{M+m} \qquad (k = 1, \dots, N) \tag{4}$$

$$\textit{Linear estimation of GCR:} \qquad \left(\frac{\|r_k\|_S}{\|r_0\|_S}\right)^{\frac{1}{k}} \leq \sqrt{1 - \left(\frac{m}{M}\right)^2} \qquad (k = 1, \dots, N) \tag{5}$$

Apart from the linear estimates, a well-known superlinear convergence estimate exists for both methods (see [6], sec. 2.2.), which uses the decomposition $S^{-1}A = I + E$. Let $s_j(E)$ $(j = 1, 2, \dots)$ denote the singular values of matrix E in decreasing order, each repeated as many times as its multiplicity. If matrix E is antisymmetric, i.e. $E^* = -E$, then the superlinear estimation in the kth iterative step is as follows:

$$\textit{Superlinear estimation of CGN:} \qquad \left(\frac{\|r_k\|_S}{\|r_0\|_S}\right)^{\frac{1}{k}} \leq \frac{2\|A^{-1}S\|_S^2}{k} \sum_{j=1}^{k} s_j^2(E) \qquad (k = 1, \dots, N) \tag{6}$$

$$\textit{Superlinear estimation of GCR:} \qquad \left(\frac{\|r_k\|_S}{\|r_0\|_S}\right)^{\frac{1}{k}} \leq \frac{\|A^{-1}S\|_S}{k} \sum_{j=1}^{k} s_j(E) \qquad (k = 1, \dots, N) \tag{7}$$

3 Numerical comparison of the iterative methods

Let us consider the following convection-diffusion type nonsymmetric elliptic problems as a special case of (1), which now depend on the constants $\varepsilon > 0$ and $\rho > 0$:

$$\begin{cases} -\varepsilon\Delta u + \rho \mathbf{w}_0 \cdot \nabla u = 1 \\ u|_{\partial\Omega} = 0 \end{cases} \tag{8}$$

We conduct numerical experiments with two vector fields used also in [7]:

(i) $\mathbf{w}_0 := (1,\ 0)$ is a constant vector field. It is used for its simplicity, and it makes the calculations more manageable.

(ii) $\mathbf{w}_0 := (-y + \frac{1}{2},\ x - \frac{1}{2})$ is a rotating vector field. This is also called an affine version of the Molenkamp–Crowley advection problem that is widely used in simplified advection models for testing numerical methods.

We are interested in which iterative method solves the resulting discretized problem in less iterative steps for different coefficients ε and ρ. For that, we fix ε, start increasing ρ from 0, and plot the number of iterative steps until convergence when TOL $:= 10^{-10}$ and $n := 50$. We do the same experiments for the three discretization methods separately, and interpret the results.

The plots for the FDM and FEM discretization are seen in figure 1 and figure 2. We analyze the plots of these methods simultaneously because they are very similar to each other. When $\rho = 0$, both methods converge instantly in one step. As ρ is increased, the number of required iterative steps increases as well. For the preconditioned CGN method, this increment seems linear with respect to ρ, while the preconditioned GCR method has a concave curve. This results in an intersection point of the two curves. Before the intersection, the preconditioned CGN method converges in less iterative steps, while afterwards this relation flips over, and the preconditioned GCR method becomes faster. For a fixed vector field, we can see that in order to reach a fixed number of iterative steps, a ten times larger ε needs an approximately ten times larger ρ. Consequently, for different ε values, we get the rescaled versions of the same graph. If we compare the graphs of the constant and rotating vector fields, the shape of the curves are very similar, only the scales of the axes differ.

The plots for the SDFEM with a usual choice of $\delta := h = \frac{1}{51}$ are seen in figure 3, while the plots with a smaller value $\delta := 5 \cdot 10^{-5}$ for the constant case can be seen in figure 4. When $\rho = 0$, both methods converge instantly. When $\delta = h$, the preconditioned CGN method performs better everywhere. For the constant vector field, there is a sudden increase in the number of iterative steps as we start increasing ρ from 0, but afterwards it rapidly approaches zero. When we take one tenth of the value of ε, the graphs look similar, but the peaks of the curves rise twice as high and get twice as close to the y-axis. For the rotating vector field, the plot for $\varepsilon = 0.001$ has similar characteristics, but the number of iterative steps settles higher than 0 as ρ gets large. For a larger ε, the curves show a different, rather increasing tendency up to $\rho = 50$.

When $\delta = 5 \cdot 10^{-5}$, mostly the preconditioned CGN method converges faster, but for a sufficiently small ε, e.g. $\varepsilon = 0.01$ or $\varepsilon = 0.001$, there is an interval around the peak where the preconditioned GCR method performs better. The scale of the axes is larger compared to the case $\delta = h$, but the shape of the graphs is similar.

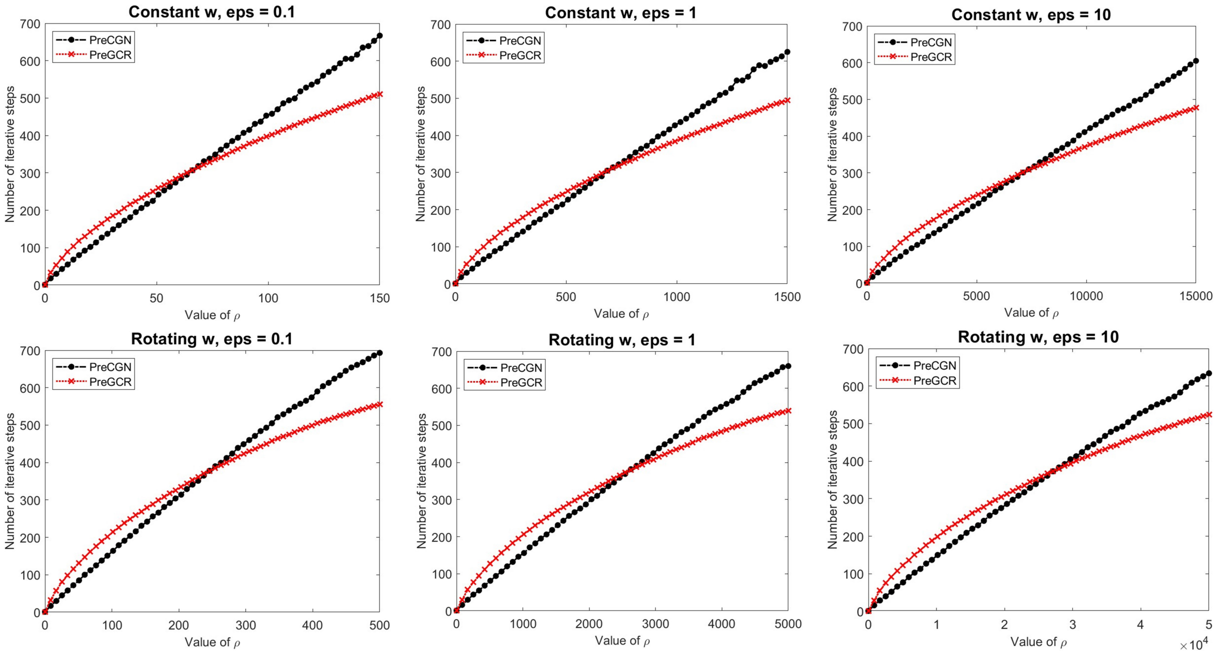

Figure 1: Number of iterative steps taken by the preconditioned CGN and GCR methods until convergence with FDM discretization when $\varepsilon = 0.1, 1, 10$ is fixed and ρ is varied.

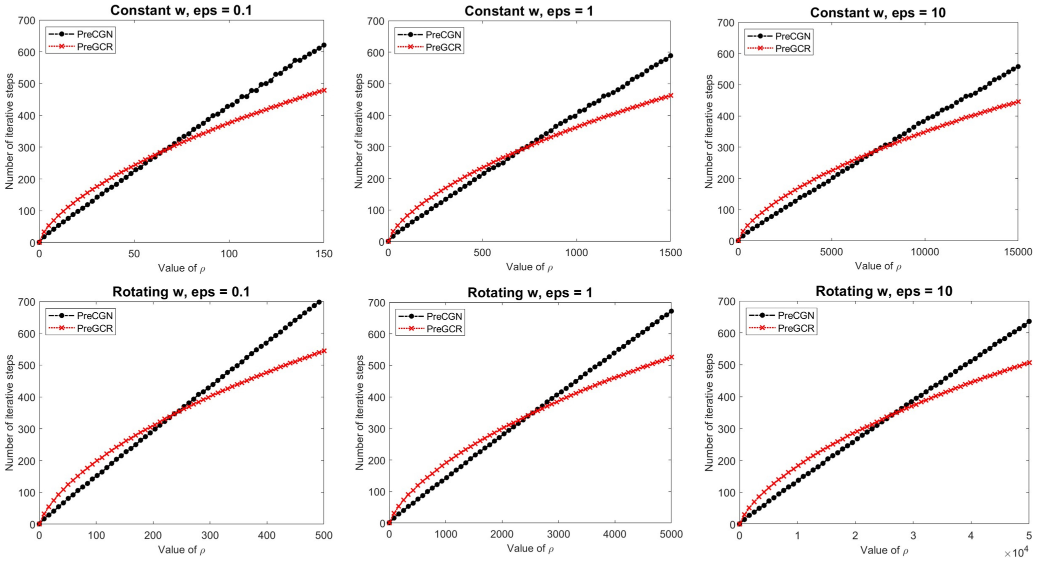

Figure 2: Number of iterative steps taken by the preconditioned CGN and GCR methods until convergence with standard FEM discretization when $\varepsilon = 0.1, 1, 10$ is fixed and ρ is varied.

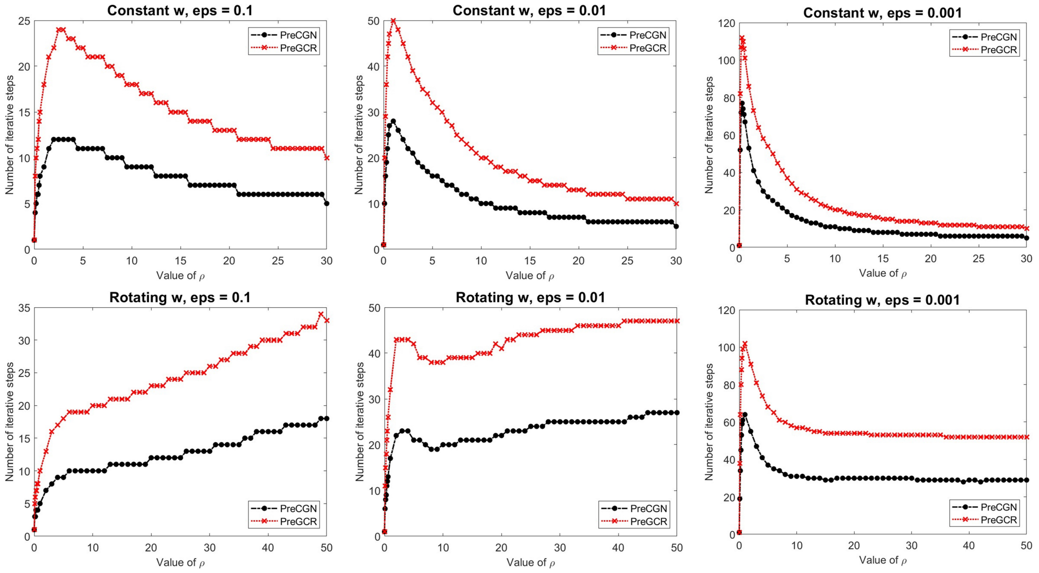

Figure 3: Number of iterative steps taken by the preconditioned CGN and GCR methods until convergence with SDFEM discretization using parameter $\delta = h$ when ε is fixed and ρ is varied.

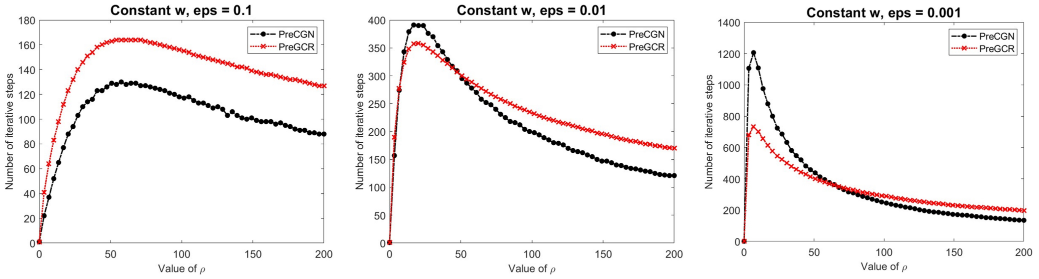

Figure 4: Number of iterative steps taken by the preconditioned CGN and GCR methods until convergence with SDFEM discretization using $\delta = 5 \cdot 10^{-5}$ when ε is fixed and ρ is varied.

4 Comparison of the linear convergence estimates

Theorem 4.1. *Let $k \in \{1,\dots,N\}$ be an arbitrary index. The linear estimation of the GCR method in the kth iterative step is better than that of the CGN method, that is, the upper bound for the convergence rate in (5) is lower than the one in (4) if and only if $\frac{M}{m} > L_k$, where m and M are defined as in (3), and L_k is the unique real root of function*

$$f_k(x) = (1-4^{\frac{1}{k}})x^3 + (3+4^{\frac{1}{k}})x^2 + 3x + 1,$$

which can be calculated as

$$L_k = \frac{c^{\frac{2}{3}}(\sqrt[3]{z-t} + \sqrt[3]{-z-t}) - 3 - c^2}{3(1-c^2)},$$

where $c = 2^{\frac{1}{k}}$, $t = 27 + 36c^2 + c^4$ and $z = (c^2-1)\sqrt{27(c^2+27)}$.

k	1	2	3	4	5	6	7	8	9	10
L_k	2.7423	5.5708	8.4388	11.3158	14.1962	17.0783	19.9614	22.8450	25.7291	28.6134

Table 1: The first ten values of L_k.

Proposition 4.2. *The finite sequence $\{L_k\}_{k=1}^{10^4}$ is strictly monotonic increasing.*

Proposition 4.3. *If the linear estimation of CGN is better than that of GCR for an index $k' \in \{1,\dots,N\}$, then it is better for any $k \in \{k',\dots,N\}$ as well.*

Proposition 4.4. *If S is the symmetric part of A, then the coercivity bound of $S^{-1}A$ is $m = 1$.*

Corollary 4.5. *The linear estimation of the CGN method in the kth iterative step is better than that of the GCR method if and only if $\|S^{-1}A\|_S < L_k$.*

Proposition 4.6. *When using standard FEM discretization, and parameter ρ in (8) satisfies*

$$\rho < \sqrt{2}\pi(L_1 - 1)\frac{\varepsilon}{\|w_0\|_{L^\infty}} \approx 7.741 \cdot \frac{\varepsilon}{\|w_0\|_{L^\infty}}, \tag{9}$$

then the linear estimation of CGN is better for every iterative step.

Proposition 4.7. *When using SDFEM discretization, and parameter ρ satisfies (9), then the linear estimation of CGN is better for every iterative step.*

Remark 4.8.

(i) By the previous calculations, we can give the following lower and upper bounds for $\|S^{-1}A\|_S$:

$$1 \le \|S^{-1}A\|_S \le 1 + \frac{\rho\|\mathbf{w}_0\|_{L^\infty}}{\varepsilon\sqrt{2}\pi}$$

By the Squeeze Theorem, $\|S^{-1}A\|_S \to 1$ as $\rho \to 0$.

(ii) When $\rho = 0$, matrix A is symmetric, therefore $S = A$ and $\|S^{-1}A\|_S = \|I\|_S = 1$.

Proposition 4.9. *When using SDFEM discretization, and parameter ρ satisfies*

$$\rho > \frac{1}{L_1 - 1} \cdot \frac{C_{w_0}}{\delta} \approx 0.574 \cdot \frac{C_{w_0}}{\delta}, \tag{10}$$

where C_{w_0} is the constant in the streamline Poincaré–Friedrichs inequality, then the linear estimation of CGN is better for every iterative step.

Remark 4.10. By the previous calculations, we can give the following lower and upper bounds for $||S^{-1}A||_S$:

$$1 \leq ||S^{-1}A||_S \leq 1 + \frac{C_{\mathbf{w}_0}}{\rho\delta}$$

By the Squeeze Theorem, $||S^{-1}A||_S \to 1$ as $\rho \to \infty$.

Corollary 4.11. *When using SDFEM discretization, if*

$$\rho < \sqrt{2}\pi(L_1 - 1)\frac{\varepsilon}{||w_0||_{L^\infty}} \quad \textit{or} \quad \rho > \frac{C_{w_0}}{(L_1 - 1)\delta}, \tag{11}$$

then the linear estimation of CGN is better for every iterative step.

Proposition 4.12. *When using SDFEM discretization with a parameter δ for which*

$$\delta > \frac{C_{w_0}||w_0||_{L^\infty}}{\sqrt{2}\pi(L_1 - 1)^2\varepsilon} \approx 0.0741 \cdot \frac{C_{w_0}||w_0||_{L^\infty}}{\varepsilon}, \tag{12}$$

then the linear estimation of CGN is better for every iterative step for any $\rho > 0$.

4.1 Application of the linear results to interpret the numerical experiments

Proposition 4.13. *For any nonsymmetric elliptic problem of the form (8), if $\rho = 0$, then both iterative methods converge after one step.*

Observation 1. *When using FEM discretization, the preconditioned CGN method is always better than the preconditioned GCR method when ρ is close to 0.*

Explanation. In Remark 4.8 we showed that $||S^{-1}A||_S \to 1$ as $\rho \to 0$, and $S^{-1}A = I$ if $\rho = 0$. The numerical tests show that whenever $||S^{-1}A||_S$ is close to 1, $||r_n||_S$ is approximated very accurately by the linear estimation. Therefore, it is reasonable to use Proposition 4.6 in case of FEM discretization. This states that if ρ is close to 0, then the linear estimation of CGN is better for every iterative step, so $||r_n||_S$ decreases more rapidly when using the CGN method. ■

Observation 2. *When using FEM discretization, the preconditioned GCR method is always better than the preconditioned CGN method when ρ is sufficiently large.*

Explanation. As ρ gets larger, $||S^{-1}A||_S$ gets larger as well. This gives advantage to the GCR method according to Corollary 4.5. On the other hand, the linear estimations are only loose approximations of the residual norms for large values of $||S^{-1}A||_S$, so this is not a sufficient explanation by itself. ■

Observation 3. *When using SDFEM discretization, the preconditioned CGN method is always better than the preconditioned GCR method when ρ is close to 0.*

Explanation. We use the argument in the explanation of Observation 1 with Proposition 4.7 in case of SDFEM. ■

Observation 4. *When using SDFEM discretization for a nonsymmetric elliptic problem where the streamline Poincaré–Friedrichs inequality is applicable (e.g. for constant vector fields), then the preconditioned CGN method is always better than the preconditioned GCR method when ρ is sufficiently large.*

Explanation. In Remark 4.10 we showed that $\|S^{-1}A\|_S \to 1$ as $\rho \to \infty$. Therefore, we can use the previous argument with Proposition 4.9. ■

Observation 5. *When using SDFEM discretization for a nonsymmetric elliptic problem where the streamline Poincaré–Friedrichs inequality is applicable, then the preconditioned CGN method is always better than the preconditioned GCR method for any $\rho > 0$ when δ is sufficiently large.*

Explanation. We use Proposition 4.12. ■

For the constant vector field $\mathbf{w}_0 = (1,\ 0)$, we can calculate the intervals where the preconditioned CGN method is always a better choice. For that, we need to calculate $\|\mathbf{w}_0\|_{L^\infty}$ and $C_{\mathbf{w}_0}$.

$\|\mathbf{w}_0\|_{L^\infty}$ is the lowest upper bound for $|\mathbf{w}_0|$ = $\sqrt{1^2+0^2} = 1$, therefore $\|\mathbf{w}_0\|_{L^\infty} = 1$.

The method for obtaining $C_{\mathbf{w}_0}$ is described in [3] (Sec. 3.2.). The characteristic curves of the vector field $\mathbf{w}_0$ can be parameterized with $\gamma_s(t) := (t,s)$, where $(s,t) \in [0,1]^2 =: K$. The absolute value of the determinant of its Jacobian matrix can be calculated as follows:

$$J_{\mathbf{w}_0}(s,t) = \left|\det\begin{pmatrix}\partial_s(t) & \partial_t(t)\\ \partial_s(s) & \partial_t(s)\end{pmatrix}\right| = \left|\det\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\right| = |-1| = 1$$

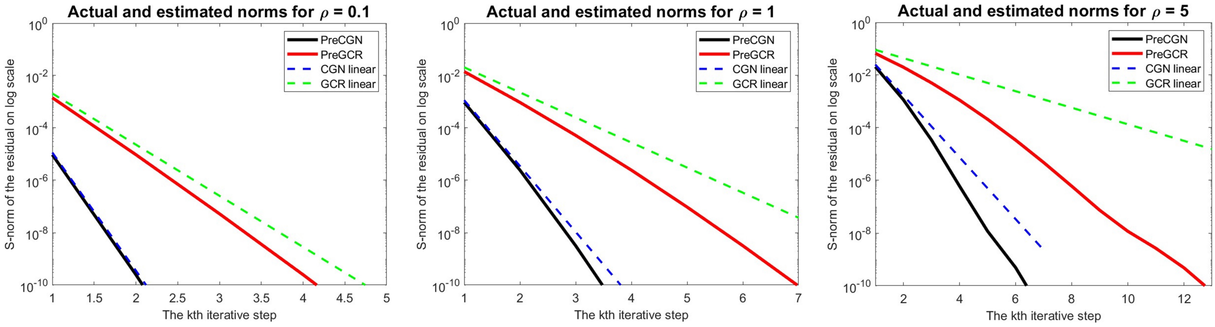

Figure 5: The actual and linear estimations of $\|r_k\|_S$ with FEM, $n = 20$, TOL $= 10^{-10}$, $\varepsilon = 1$, $\mathbf{w}_0 = (1,\ 0)$ and $\rho = 0.1, 1, 5$.

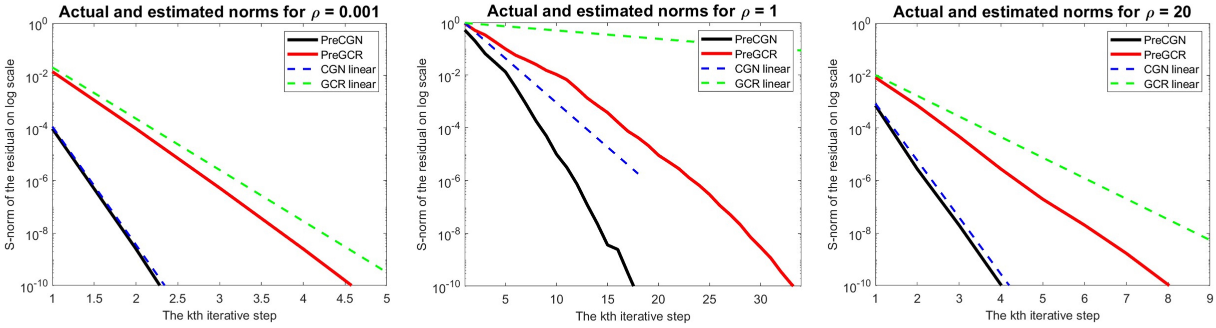

Figure 6: The actual and linear estimations of $\|r_k\|_S$ with SDFEM, $n = 20$, TOL $= 10^{-10}$, $\delta = h$, $\varepsilon = 0.01$, $\mathbf{w}_0 = (1,\ 0)$ and $\rho = 0.001, 1, 20$.

This is bounded from below and above by $\mu = \tilde{\mu} := 1$. From this, we can calculate

$$C_{\mathbf{w}_0} = \operatorname{diam}(K) \cdot \sqrt{\tilde{\mu}/\mu} = \operatorname{diam}(K) = \sqrt{1^2 + 1^2} = \sqrt{2}.$$

According to the relation in (9), the CGN method is better in the interval $(0, 7.741 \cdot \varepsilon)$ when the discretization is executed with the FEM, which is the interval $(0, 7.741)$ for $\varepsilon = 1$.

According to the relation in (11), the CGN method is better in the domain $\mathbb{R}^+ \setminus \left[7.741 \cdot \varepsilon,\ \frac{0.812}{\delta}\right]$ when the discretization is executed with the SDFEM, which is $\mathbb{R}^+ \setminus [0.0774,\ 16.917]$ for $\varepsilon = 0.01$ and $\delta = 0.048$.

The plot of the residual norms $\|r_k\|_S$ and the corresponding linear estimations can be seen in figure 5 for $\varepsilon = 1$ with FEM and in figure 6 for $\varepsilon = 0.01$ with SDFEM. These graphs confirm that according to our calculations if $\rho \in (0, 7.741)$ in figure 5 and $\rho \in \mathbb{R}^+ \setminus [0.0774,\ 16.917]$ in figure 6, then the linear estimations follow closely the residual norms, and the linear estimation of CGN is always below that of GCR. We can also see that the linear estimations in figure 5 start deteriorating as ρ is increased, and in case of SDFEM in figure 6, the linear estimations do not approximate well when $\rho = 1$, which is in the excluded interval.

5 Comparison of the superlinear convergence estimates

Apart from the previously analyzed cases, numerical experiments show that the linear estimation of the residual norm does not reflect the actual convergence. The reason for this is that after a short linear phase, the rate of convergence speeds up and gets into the superlinear phase. Thus for further analysis, we need to utilize the superlinear estimates introduced in Section 2.3.

Proposition 5.1. *Let $\|A^{-1}S\|_S = 1$ be satisfied, and suppose that E is antisymmetric. The superlinear estimation of GCR in the kth iterative step is better than that of CGN exactly when*

$$\sum_{j=1}^{k} s_j(E) < 2 \sum_{j=1}^{k} s_j^2(E). \tag{13}$$

Proposition 5.2. *When using SDFEM discretization, if $\varepsilon = 0$ and $\mathbf{w}_0 = (1,\ 0)$, then the eigenvalues of operator E are*

$$\mu_j(E) = \frac{i}{2\pi\rho\delta j} \quad (j \in \mathbb{Z} \setminus \{0\}). \tag{14}$$

Proposition 5.3. *When using SDFEM discretization, if $\varepsilon = 0$ and $\mathbf{w}_0 = (1,\ 0)$, then*

$$s_{2j}(E) = s_{2j-1}(E) = \frac{1}{2\pi\rho\delta j} \quad (j = 1, 2, \ldots).$$

Proposition 5.4. *Let $k \in \mathbb{N}$ be arbitrary and $k' := 2k$. When using SDFEM discretization, if $\varepsilon = 0$ and $\mathbf{w}_0 = (1,\ 0)$, then the superlinear estimation of the GCR method in the k'-th iterative step is better than that of the CGN method if*

$$\rho < \frac{1}{\pi\delta} \frac{\sum\limits_{j=1}^{k} \frac{1}{j^2}}{\sum\limits_{j=1}^{k} \frac{1}{j}} \approx \frac{1}{\pi\delta} \frac{\frac{\pi^2}{6} - \frac{1}{k}}{0.5772 + \ln k + \frac{1}{2k}}. \tag{15}$$

Proposition 5.5. *Let $k \in \mathbb{N}$ be an arbitrary index and $E_0 := \frac{1}{\rho}E$. When using FEM discretization, the superlinear estimation of the GCR method in the kth iterative step is better than that of the CGN method if*

$$\rho > \frac{\sum_{j=1}^{k} s_j(E_0)}{2\sum_{j=1}^{k} s_j^2(E_0)}. \tag{16}$$

5.1 Limitations of using the linear and superlinear estimates

So far, we have explained why the CGN method is better on the left side of the FEM plots in figure 2 using Observation 1, and with a similar argument we have shown why the CGN method is better on the left and right sides of the SDFEM plots in figure 3 and 4 using Observation 3 and 4. What remain to be answered is that why the GCR method is better on the right side of the FEM plots, and what determines whether the CGN or GCR method is better near the peak on the left side of the SDFEM plots. These have been partially explained in Observation 2 and Proposition 5.5 in case of FEM and in Observation 5 and Proposition 5.4 in case of SDFEM.

Unfortunately, neither the comparison of the linear nor the superlinear estimates can be used directly to explain these outcomes. The reason for this is that none of these reflect the actual convergence of the residual norms. This is illustrated in figure 7 and 8.

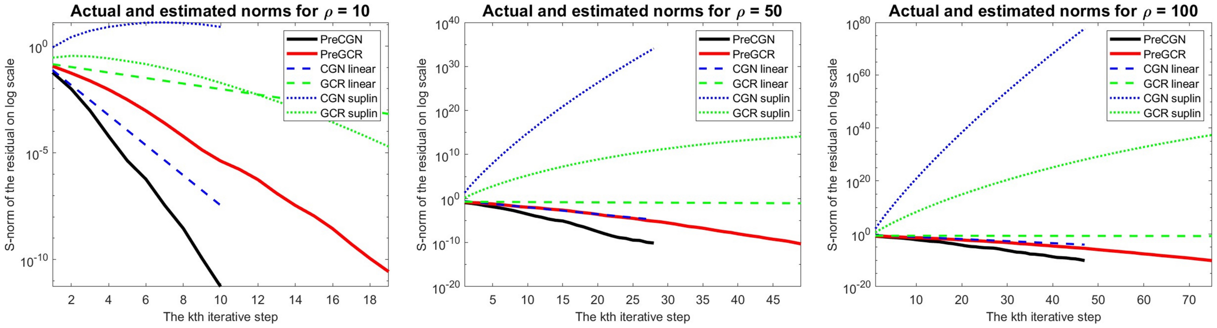

Figure 7: The actual residual norms and their linear and superlinear estimations with FEM, $n = 20$, TOL $= 10^{-10}$, $\varepsilon = 1$, $\mathbf{w}_0 = (1,\ 0)$ and $\rho = 10, 50, 100$.

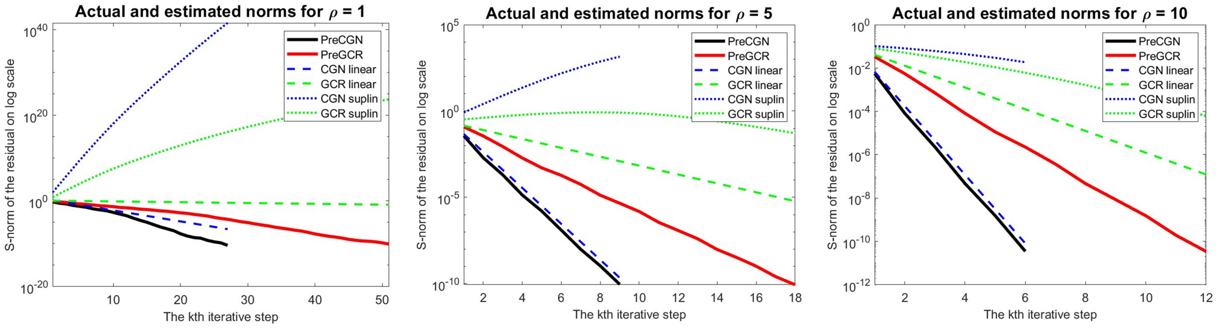

Figure 8: The actual residual norms and their linear and superlinear estimations with SDFEM, $n = 20$, TOL $= 10^{-10}$, $\delta = h$, $\varepsilon = 0$, $\mathbf{w}_0 = (1,\ 0)$ and $\rho = 1, 5, 10$.

References

[1] NACHTIGAL, N. M.; REDDY, S. C.; TREFETHEN, L. N.: *How fast are nonsymmetric matrix iterations?*, SIAM J. Matrix Anal. Appl. Vol. 13, No. 3, pp. 778-795, July 1992.

[2] KARÁTSON, J.; HORVÁTH, R.: *Numerical Methods for Elliptic Partial Differential Equations.* URL: `https://kajkaat.web.elte.hu/pdnmell-ang-2022.pdf`

[3] AXELSSON, O.; KARÁTSON, J.; KOVÁCS, B.: *Robust Preconditioning Estimates for Convection-Dominated Elliptic Problems via a Streamline Poincaré–Friedrichs Inequality.* SIAM Journal on Numerical Analysis, Vol. 52, Iss. 6, 2014.

[4] BAKOS, I.: *Konvekció-diffúziós egyenletek.* BSc thesis, BME, 2014.

[5] SAAD, Y.: *Iterative methods for sparse linear systems.* SIAM, Philadelphia, 2003.

[6] AXELSSON, O.; KARÁTSON, J.; MAGOULÈS, F.: *Robust Superlinear Krylov Convergence for Complex Noncoercive Compact-Equivalent Operator Preconditioners.* SIAM Journal on Numerical Analysis, Vol. 61, Iss. 2, 2023.

[7] KARÁTSON, J.: *Superlinear Krylov convergence under streamline diffusion FEM for convection-dominated elliptic operators.* Numer Linear Algebra Appl. 2024;e2586.

[8] SAITO, T.; ISHIWATA, E.; HASEGAWA, H.: *Analysis of the GCR method with mixed precision arithmetic using QuPAT.* Journal of Computational Science, Vol. 3, Issue 3, pp. 87-91, 2012.

[9] THE MATHWORKS, INC.: *MATLAB R2024a Update 6.* Natick, Massachusetts, 2024. URL: `https://www.mathworks.com/help/matlab/`