

EÖTVÖS LORÁND UNIVERSITY
FACULTY OF SCIENCE

Comparison of iterative methods for discretized nonsymmetric elliptic problems

TDK thesis

Lados Bálint István

E-MAIL: ladosbalint01@gmail.com

Applied Mathematics MSc (2nd Year)
Applied Analysis specialization

SUPERVISOR:

Karátson János

Professor
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).

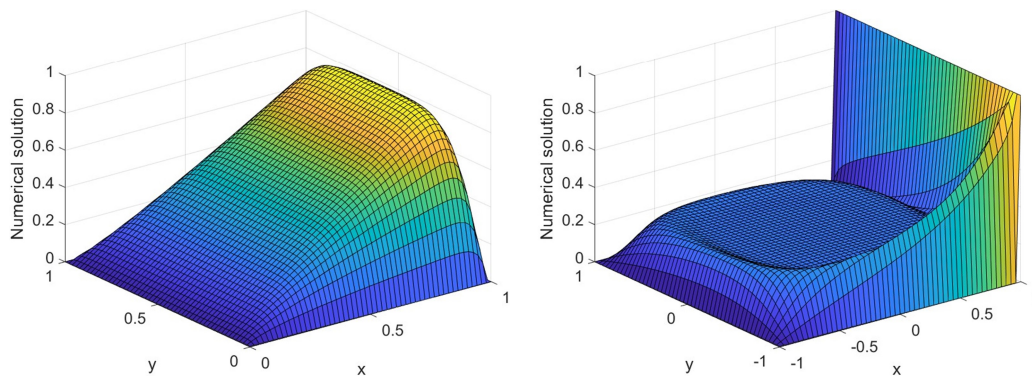


Figure 1: Numerical solution of two convection-dominated problems with a constant vector field on the left, and a recirculating wind on the right.

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 results are then used to explain the outcome of the numerical tests, and along with that the limitations of this approach are discussed as well.

2 Preliminaries

2.1 Nonsymmetric elliptic problems

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

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

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 $\text{div}(\mathbf{w}) = 0$;
- (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

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

We perform one of the discretization methods on the problem in (1) to give a finite dimensional approximation. The three discretization methods (FDM, standard FEM and SDFEM) have been implemented in MATLAB ([9]) with the constructions discussed in Section 2.2, 2.3 and 2.4.

2.2 Finite Difference Method (FDM)

Let us create an equidistant mesh on the unit square with $(n+2)^2$ grid points such that the distance is $h := \frac{1}{n+1}$ and the grid points are

$$w_h := \{(ih, jh) : i, j = 0, 1, \dots, n+1\}.$$

We introduce the variables $u_{i,j}$ ($i, j = 0, 1, \dots, n+1$) which correspond to the numerical solution at point (ih, jh) . We use Dirichlet boundary condition, so the variables corresponding to the grid points on the boundary ($i \in \{0, n+1\}$ or $j \in \{0, n+1\}$) can be set to 0. For the inner grid points, we use the central finite difference scheme:

$$\begin{aligned} \partial_x u_{i,j} &\approx \frac{u_{i+1,j} - u_{i-1,j}}{2h} & \partial_y u_{i,j} &\approx \frac{u_{i,j+1} - u_{i,j-1}}{2h} \\ \partial_{xx} u_{i,j} &\approx \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} & \partial_{yy} u_{i,j} &\approx \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h^2} \end{aligned}$$

The discretization results in a system of linear equations with $N := n^2$ unknowns.

2.3 Finite Element Method (FEM)

This discretization method is based on the weak form of the boundary value problem, which was already specified in (2). Let us consider the equidistant triangulation of the unit square with the same grid points as before. We can define the finite dimensional subspace $V_h \subset H_0^1(\Omega)$ consisting of functions defined on Ω that are linear on each triangle of the triangulation (also called Courant elements). With this construction, V_h is N dimensional, and we consider the basis functions $\phi_1, \dots, \phi_N$ that are 0 at all the grid points except for one point, where it is 1.

We look for the numerical solution $u_h \in V_h$ as the linear combination of the basis functions:

$$u_h = \sum_{j=1}^N c_j \phi_j. \quad (3)$$

In the standard version of FEM, we substitute u_h into the weak form with any basis function ϕ_i :

$$\int_{\Omega} (\alpha \nabla u_h \cdot \nabla \phi_i + (\mathbf{w} \cdot \nabla u_h) \phi_i) = \sum_{j=1}^N c_j \int_{\Omega} (\alpha \nabla \phi_j \cdot \nabla \phi_i + (\mathbf{w} \cdot \nabla \phi_j) \phi_i) =: \sum_{j=1}^N c_j a_{ij}$$

If we consider matrix A with components a_{ij} and we set $b_i := \int_{\Omega} f \phi_i$ ($i = 1, \dots, N$), then we get a system of linear equations $Ac = b$ for the coefficients of u_h in (3).

2.4 Streamline Diffusion Finite Element Method (SDFEM)

This can be considered as an extension of the previously introduced finite element method. It 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 $\delta = O(h)$.

The components of matrix A are $a_{ij} := \int_{\Omega} (\alpha \nabla \phi_j \cdot \nabla \phi_i + (\mathbf{w} \cdot \nabla \phi_j) \phi_i + \delta (\mathbf{w} \cdot \nabla \phi_j)(\mathbf{w} \cdot \nabla \phi_i))$, and vector b is modified as well, so that $b_i := \int_{\Omega} f (\phi_i + \delta \mathbf{w} \cdot \nabla \phi_i)$.

If we chose $\delta := 0$, we would get back the standard FEM.

With our choice of an equidistant triangulation with Courant elements, it turns out that $\int_{\Omega} \partial_x \phi_i = \int_{\Omega} \partial_y \phi_i = 0$, and the approximation of b_i with the one-point quadrature simplifies to the standard form:

$$b_i = \int_{\Omega} f(\phi_i + \delta \mathbf{w} \cdot \nabla \phi_i) = \int_{\Omega} f \phi_i + \delta \int_{\Omega} f \mathbf{w} \cdot \nabla \phi_i \approx f(x_i, y_i) \int_{\Omega} \phi_i = f(x_i, y_i) \cdot h^2.$$

If $\mathbf{w}$ is a constant vector field, the integrals $\int_{\Omega} (\mathbf{w} \cdot \nabla \phi_j) \phi_i$ and $\int_{\Omega} \delta (\mathbf{w} \cdot \nabla \phi_j)(\mathbf{w} \cdot \nabla \phi_i)$ in a_{ij} can be calculated as in [4], ch. 5. If $\mathbf{w} = (w_x, w_y)$ is not constant, we are still able to reduce the problem to the constant case using a one-point quadrature and without calculating more integrals of the product of the basis functions and their derivatives, as follows:

$$\int_{\Omega} (\mathbf{w} \cdot \nabla \phi_j) \phi_i \approx w_x(x_i, y_j) \int_{\Omega} ((1, 0) \cdot \nabla \phi_j) \phi_i + w_y(x_i, y_j) \int_{\Omega} ((0, 1) \cdot \nabla \phi_j) \phi_i$$

$$\begin{aligned} \delta \int_{\Omega} (\mathbf{w} \cdot \nabla \phi_j)(\mathbf{w} \cdot \nabla \phi_i) &\approx \delta w_x^2(x_i, y_j) \int_{\Omega} ((1, 0) \cdot \nabla \phi_j)((1, 0) \cdot \nabla \phi_i) + \\ + \delta w_y^2(x_i, y_j) &\int_{\Omega} ((0, 1) \cdot \nabla \phi_j)((0, 1) \cdot \nabla \phi_i) + \delta w_x(x_i, y_j) w_y(x_i, y_j) \int_{\Omega} (((1, 1) \cdot \nabla \phi_j)((1, 1) \cdot \nabla \phi_i) - \\ &- ((1, 0) \cdot \nabla \phi_j)((1, 0) \cdot \nabla \phi_i) - ((0, 1) \cdot \nabla \phi_j)((0, 1) \cdot \nabla \phi_i)) \end{aligned}$$

2.5 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

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

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. In the algorithms, instead of calculating vectors in the form of $x = S^{-1}y$, we consider the corresponding auxiliary equations $Sx = y$, which can be solved by the built-in solver $x = S \backslash y$ in MATLAB.

ALGORITHM 1. The Preconditioned CGN

```

 $u_0 := \mathbf{0};$ 
 $r_0 := S^{-1}Au_0 - S^{-1}b;$ 
 $s_0 := S^{-1}A^T r_0;$ 
 $p_0 := s_0;$ 
 $n := 0;$ 
while  $\|r_n\|_S > TOL$  do
     $z_n := S^{-1}Ap_n;$ 
     $\alpha_n = -\frac{\|s_n\|_S^2}{\|z_n\|_S^2};$ 
     $u_{n+1} := u_n + \alpha_n p_n;$ 
     $r_{n+1} := r_n + \alpha_n z_n;$ 
     $s_{n+1} := S^{-1}A^T r_{n+1};$ 
     $\beta_n = \frac{\|s_{n+1}\|_S^2}{\|s_n\|_S^2};$ 
     $p_{n+1} := s_{n+1} + \beta_n p_n;$ 
     $n := n + 1;$ 
end
```

ALGORITHM 2. The Preconditioned GCR

```

 $u_0 := \mathbf{0};$ 
 $r_0 := S^{-1}b - S^{-1}Au_0;$ 
 $p_0 := r_0;$ 
 $z_0 := S^{-1}Ap_0;$ 
 $n := 0;$ 
while  $\|r_n\|_S > TOL$  do
     $\alpha_n := \frac{\langle r_n, z_n \rangle_S}{\|z_n\|_S^2};$ 
     $u_{n+1} := u_n + \alpha_n p_n;$ 
     $r_{n+1} := r_n - \alpha_n z_n;$ 
     $s_n := S^{-1}Ar_{n+1};$ 
    for  $i = 0, 1, \dots, n$  do
         $\beta_{i,n} := -\frac{\langle s_n, z_i \rangle_S}{\|z_i\|_S^2};$ 
    end
     $p_{n+1} := r_{n+1} + \sum_{i=0}^n \beta_{i,n} p_i;$ 
     $z_{n+1} := s_n + \sum_{i=0}^n \beta_{i,n} z_i;$ 
     $n := n + 1;$ 
end
```

2.6 Linear and superlinear convergence estimates

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

$$\begin{aligned} m &:= \inf \{ \langle S^{-1}Ac, c \rangle_S : \|c\|_S = 1 \} = \inf \{ \langle Ac, c \rangle : \|c\|_S = 1 \} > 0 \\ M &:= \|S^{-1}A\|_S = \sup \{ \langle Ac, d \rangle : \|c\|_S = \|d\|_S = 1 \} \end{aligned} \quad (4)$$

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

$$\text{Linear estimation of CGN:} \quad \left(\frac{\|r_k\|_S}{\|r_0\|_S} \right)^{\frac{1}{k}} \leq 2^{\frac{1}{k}} \frac{M-m}{M+m} \quad (k = 1, \dots, N) \quad (5)$$

$$\text{Linear estimation of GCR:} \quad \left(\frac{\|r_k\|_S}{\|r_0\|_S} \right)^{\frac{1}{k}} \leq \sqrt{1 - \left(\frac{m}{M} \right)^2} \quad (k = 1, \dots, N) \quad (6)$$

Apart from the linear estimates, a well-known superlinear convergence estimate exists for both methods (see [6], sec. 2.2.), which uses the decomposition

$$S^{-1}A = I + E.$$

If $\mu_j(C)$ and $s_j(C)$ ($j = 1, 2, \dots$) denote the eigenvalues and singular values of a matrix C in decreasing order, each repeated as many times as its multiplicity, then the superlinear bound in the k th iterative step is as follows:

$$\text{Superlinear estimation of CGN:} \quad \left(\frac{\|r_k\|_S}{\|r_0\|_S} \right)^{\frac{1}{k}} \leq \frac{2\|A^{-1}S\|_S^2}{k} \sum_{j=1}^k (|\mu_j(E^* + E)| + s_j^2(E)) \quad (7)$$

$$\text{Superlinear estimation of GCR:} \quad \left(\frac{\|r_k\|_S}{\|r_0\|_S} \right)^{\frac{1}{k}} \leq \frac{\|A^{-1}S\|_S}{k} \sum_{j=1}^k s_j(E) \quad (8)$$

If matrix E is antisymmetric, i.e. $E^* = -E$, then $\mu_j(E^* + E) = 0$, and the superlinear estimation of CGN simplifies as:

$$\left(\frac{\|r_k\|_S}{\|r_0\|_S} \right)^{\frac{1}{k}} \leq \frac{2\|A^{-1}S\|_S^2}{k} \sum_{j=1}^k s_j^2(E) \quad (9)$$

An even sharper superlinear estimate exists for the GCR method, which is described e.g. in [7], that uses the product of the singular values $s_j(E)$:

$$\left(\frac{\|r_k\|_S}{\|r_0\|_S} \right)^{\frac{1}{k}} \leq \|A^{-1}S\|_S \left(\prod_{j=1}^k s_j(E) \right)^{\frac{1}{k}} \quad (10)$$

3 New results

3.1 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 $\rho > 0$:

$$\begin{cases} -\varepsilon \Delta u + \rho \mathbf{w}_0 \cdot \nabla u = 1 \\ u|_{\partial\Omega} = 0 \end{cases} \quad (11)$$

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 $\text{TOL} := 10^{-10}$ 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 2 and figure 3. 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 4, while the plots with a smaller value $\delta := 5 \cdot 10^{-5}$ for the constant case can be seen in figure 5. 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.

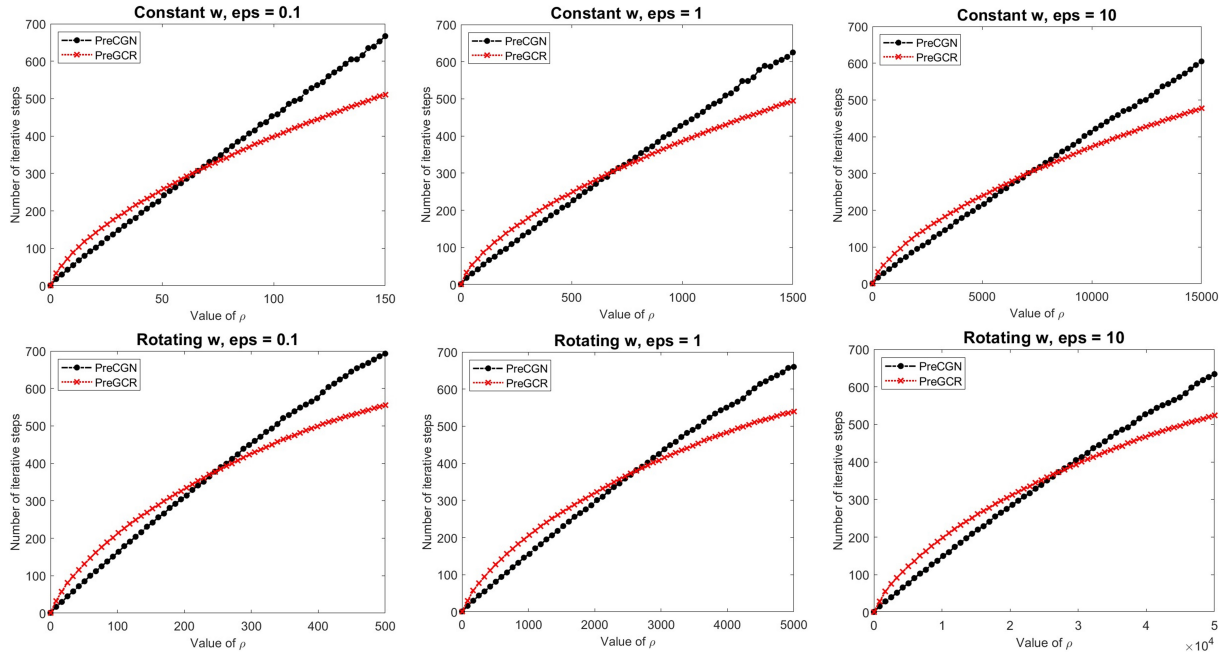


Figure 2: Number of iterative steps taken by the preconditioned CGN and GCR methods until convergence with FDM discretization when $\varepsilon = 0.1, 1, 10$ is fixed and ρ is varied.

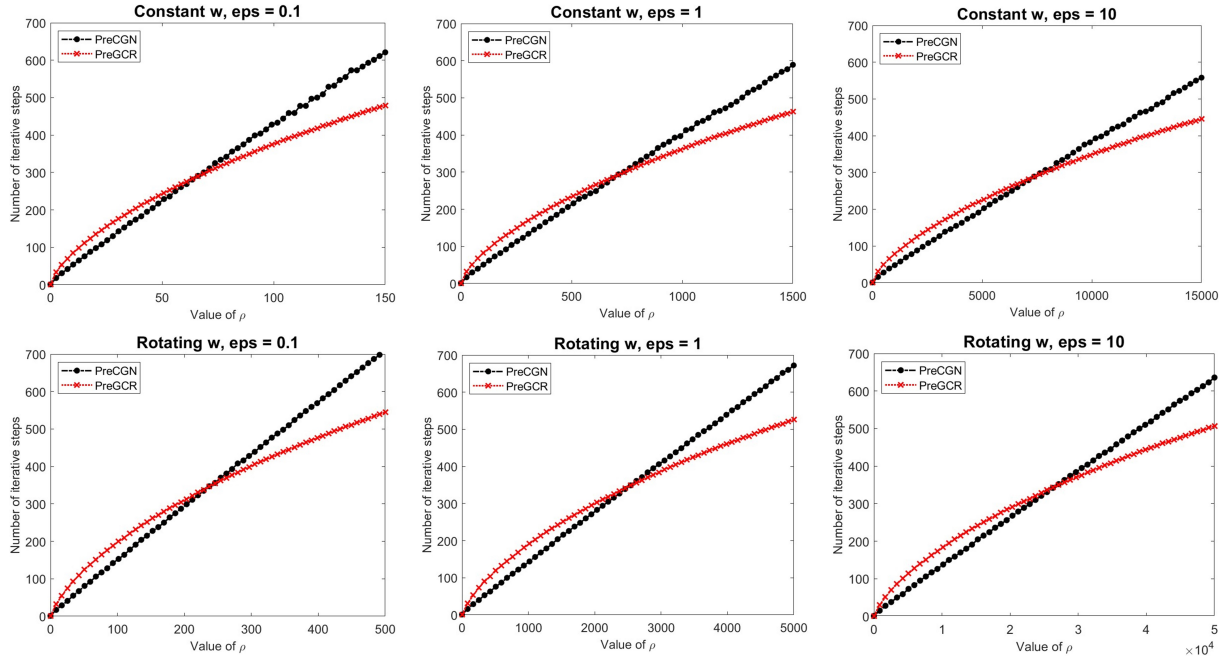


Figure 3: Number of iterative steps taken by the preconditioned CGN and GCR methods until convergence with standard FEM discretization when $\varepsilon = 0.1, 1, 10$ is fixed and ρ is varied.

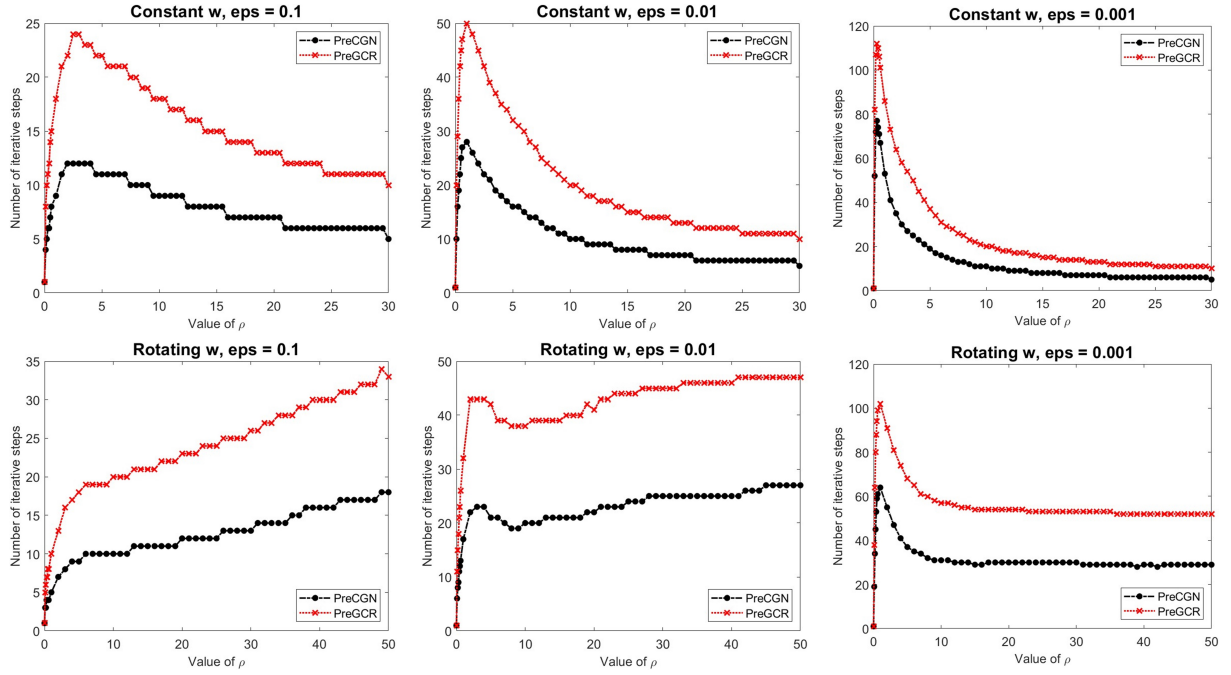


Figure 4: Number of iterative steps taken by the preconditioned CGN and GCR methods until convergence with SDFEM discretization using parameter $\delta = h$ when $\varepsilon = 0.1, 0.01, 0.001$ is fixed and ρ is varied.

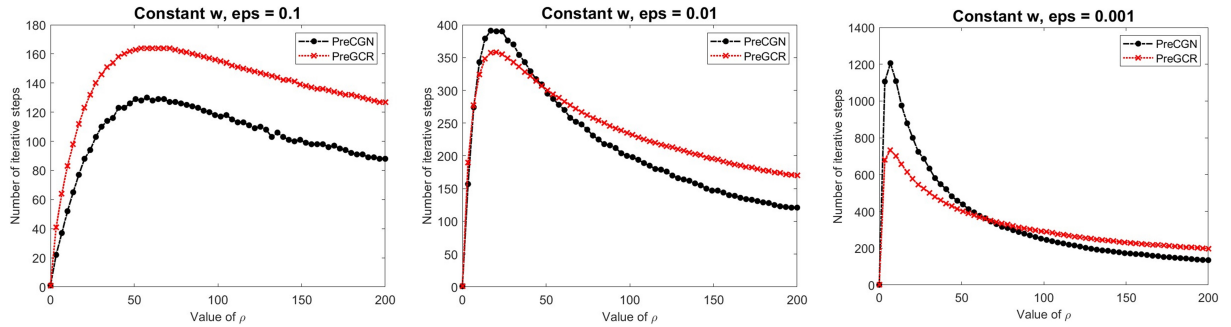


Figure 5: Number of iterative steps taken by the preconditioned CGN and GCR methods until convergence with SDFEM discretization using parameter $\delta = 5 \cdot 10^{-5}$ when $\varepsilon = 0.1, 0.01, 0.001$ is fixed and ρ is varied.

3.2 Comparison of the linear convergence estimates

Theorem 3.1. *Let $k \in \{1, \dots, N\}$ be an arbitrary index. The linear estimation of the GCR method in the k th iterative step is better than that of the CGN method, that is, the upper bound for the convergence rate in (6) is lower than the one in (5) if and only if $\frac{M}{m} > L_k$, where m and M are defined as in (4), and L_k is the unique real root of function*

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

Proof. Let $k \in \{1, \dots, N\}$ be fixed and $c := 2^{\frac{1}{k}}$. Let us rewrite the bounds in (5) and (6) using $\kappa := \frac{M}{m} \geq 1$. If $\kappa = 1$, then both estimations are zero, so we can suppose that $\kappa > 1$.

$$\begin{aligned} 2^{\frac{1}{k}} \frac{M-m}{M+m} &= c \frac{\kappa-1}{\kappa+1} = c \frac{\kappa^2-1}{(\kappa+1)^2} = c \frac{(\sqrt{\kappa^2-1})^2}{(\kappa+1)^2} \\ \sqrt{1 - \left(\frac{m}{M}\right)^2} &= \sqrt{1 - \frac{1}{\kappa^2}} = \sqrt{\frac{\kappa^2-1}{\kappa^2}} = \frac{\sqrt{\kappa^2-1}}{\kappa} \end{aligned}$$

Now we can compare the two quantities:

$$\begin{aligned} \frac{\sqrt{\kappa^2-1}}{\kappa} &< c \frac{(\sqrt{\kappa^2-1})^2}{(\kappa+1)^2} \Leftrightarrow \frac{1}{\kappa} < c \frac{\sqrt{\kappa^2-1}}{(\kappa+1)^2} \Leftrightarrow (\kappa+1)^2 < c\kappa\sqrt{\kappa^2-1} \\ \Leftrightarrow (\kappa+1)^4 &< c^2\kappa^2(\kappa-1)(\kappa+1) \Leftrightarrow f_k(\kappa) := (\kappa+1)^3 - c^2\kappa^2(\kappa-1) < 0 \end{aligned}$$

We can expand and reorganize the terms in $f_k(\kappa)$:

$$f_k(\kappa) = \kappa^3 + 3\kappa^2 + 3\kappa + 1 - c^2\kappa^3 + c^2\kappa^2 = (1 - c^2)\kappa^3 + (3 + c^2)\kappa^2 + 3\kappa + 1$$

The coefficients of $f_k(\kappa)$ are $\tilde{a} := 1 - c^2$, $\tilde{b} := 3 + c^2$, $\tilde{c} := 3$ and $\tilde{d} := 1$. The roots of $f_k(\kappa)$ can be found by reducing the cubic polynomial to the form $x^3 + px + q$, where

$$\begin{aligned} p &= \frac{\tilde{c}}{\tilde{a}} - \frac{\tilde{b}^2}{3\tilde{a}^2} = \frac{-c^2(15 + c^2)}{3(1 - c^2)^2} \\ q &= \frac{\tilde{d}}{\tilde{a}} + \frac{2\tilde{b}^3}{27\tilde{a}^3} - \frac{\tilde{b}\tilde{c}}{3\tilde{a}^2} = \frac{2c^2(27 + 36c^2 + c^4)}{27(1 - c^2)^3}. \end{aligned}$$

We can proceed by first calculating

$$\sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3} = \frac{c^2(c^2-1)\sqrt{27(c^2+27)}}{27(1-c^2)^3},$$

and then substituting into the cubic formula to get that

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

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

We can verify that L_k is the only real root of $f_k(\kappa)$ by checking the sign of its discriminant:

$$\Delta = -(4p^3 + 27q^2) < 0 \Leftrightarrow 4p^3 + 27q^2 > 0$$

$$\begin{aligned}
&\Leftrightarrow 4p^3 + 27q^2 = \frac{-4c^6(15+c^2)^3 + 4c^4(27+36c^2+c^4)^2}{27(1-c^2)^6} = \\
&= \frac{-13500c^6 - 2700c^8 - 180c^{10} - 4c^{12} + 2916c^4 + 5184c^8 + 4c^{12} + 7776c^6 + 216c^8 + 288c^{10}}{27(1-c^2)^6} = \\
&= \frac{c^4(108c^6 + 2700c^4 - 5724c^2 + 2916)}{27(1-c^2)^6} = \frac{4c^4(c^6 + 25c^4 - 53c^2 + 27)}{(1-c^2)^6} > 0
\end{aligned}$$

It is sufficient to show that $c^6 + 25c^4 - 53c^2 + 27 > 0$. We defined $c = 2^{\frac{1}{k}}$ when $k \geq 1$, so we can restrict the domain to $1 < c \leq 2$. We can substitute $x := c^2$ to show that $g(x) := x^3 + 25x^2 - 53x + 27 = (x-1)^2(x+27) > 0$, if $1 < x \leq 4$. If $1 < x$, then all the terms in the factored form of g are positive, therefore g is positive in the interval $(1, 4]$.

The leading coefficient of $f_k(\kappa)$ is $\tilde{a} = 1 - c^2$, which is negative as $1 < c$, therefore $\lim_{\kappa \rightarrow \infty} f_c(\kappa) = -\infty$ and $\lim_{\kappa \rightarrow -\infty} f_c(\kappa) = \infty$. Function $f_k(\kappa)$ is continuous, so Bolzano's Theorem is applicable, which ensures that $f_k(\kappa) < 0 \Leftrightarrow \kappa > L_k$. $\square$

Proposition 3.2. *The finite sequence $\{L_k\}_{k=1}^{10^4}$ is strictly monotonic increasing.*

Proof. I verified it by generating the sequence in MATLAB and checking that the difference of the consecutive terms is always positive. $\square$

Remark 3.3.

- (i) My conjecture is that the infinite sequence $\{L_k\}_{k=1}^{\infty}$ is strictly monotonic increasing as well. It could be proved by calculating the derivative of L_k with respect to k and checking that it is positive when $k \geq 1$. However, the derivative has a very complicated form, which makes it hard to determine its sign.
- (ii) In the numerical experiments, our grid parameter n in the discretization is at most 100, so the iterative methods converge to the exact solution in at most $N = 10^4$ steps. Proposition 3.2 ensures that $\{L_k\}$ is strictly monotonic increasing up to this point.
- (iii) The finite sequence $\{L_k\}_{k=1}^{10^6}$ can be approximated very accurately by linear regression. The least squares method gives the approximation $L_k \approx 2.8854 \cdot k - 0.25$, which has a root mean square error of $1.45 \cdot 10^{-4}$.

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 3.4. *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.*

Proof. By Theorem 3.1 we know that the estimation of CGN is better for an index k' if and only if $\frac{M}{m} < L_{k'}$. The sequence $\{L_k\}_{k=1}^N$ is strictly monotonic increasing by Proposition 3.2, therefore $k' < k$ implies that $L_{k'} < L_k$, so $\frac{M}{m} < L_k$ is satisfied for any index $k \in \{k', \dots, N\}$. $\square$

Remark 3.5. The previous propositions and Theorem 3.1 are generally true for any matrix A and S . We use only from this point that matrix A is obtained from the discretization of a nonsymmetric elliptic problem, and S is the symmetric part of A .

Proposition 3.6. *If S is the symmetric part of A , then the coercivity bound of $S^{-1}A$ is $m = 1$.*

Proof. According to (4), it is sufficient to show that $\langle Ac, c \rangle = \|c\|_S^2$ is satisfied.

$$\|c\|_S^2 = \langle Sc, c \rangle = \left\langle \left(\frac{A + A^T}{2} \right) c, c \right\rangle = \frac{1}{2} \langle Ac + A^T c, c \rangle = \langle Ac, c \rangle \quad \square$$

Corollary 3.7. *The linear estimation of the CGN method in the k th iterative step is better than that of the GCR method if and only if $\|S^{-1}A\|_S < L_k$.*

Proposition 3.8. *When using standard FEM discretization, and parameter ρ in (11) satisfies*

$$\rho < \sqrt{2}\pi(L_1 - 1) \frac{\varepsilon}{\|\mathbf{w}_0\|_{L^\infty}} \approx 7.741 \cdot \frac{\varepsilon}{\|\mathbf{w}_0\|_{L^\infty}}, \quad (12)$$

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

Proof. According to Proposition 3.4, if the estimation of CGN is already better for $k = 1$, then it is better for any other iterative step as well. Let us consider the bilinear form $a(u, v)$ corresponding to the weak form of the studied nonsymmetric elliptic problem:

$$a(u, v) := \int_{\Omega} (\varepsilon \nabla u \cdot \nabla v + (\rho \mathbf{w}_0 \cdot \nabla u) v)$$

This is bounded with the H_0^1 -norm as

$$|a(u, v)| \leq (\varepsilon + \lambda_1^{-\frac{1}{2}} \rho \|\mathbf{w}_0\|_{L^\infty}) \|u\|_{H_0^1} \|v\|_{H_0^1}, \quad \forall u, v \in H_0^1(\Omega),$$

where the constant in the Poincaré–Friedrichs inequality is $\lambda_1 = 2\pi^2$, which is the smallest eigenvalue of operator $-\Delta$ in Ω with homogeneous Dirichlet boundary condition. It is easy to see that $\|u\|_{H_0^1} = \frac{1}{\sqrt{\varepsilon}} \|u\|_S = \frac{1}{\sqrt{\varepsilon}} \sqrt{\int_{\Omega} \varepsilon |\nabla u|^2}$, so the bound of $a(u, v)$ with respect to the S -norm is

$$\tilde{M} := 1 + \frac{\rho \|\mathbf{w}_0\|_{L^\infty}}{\varepsilon \sqrt{2}\pi}. \quad (13)$$

By equivalent preconditioning, $\|S^{-1}A\|_S$ inherits the upper bound of $a(u, v)$ owing to (4). That is, $\|S^{-1}A\|_S \leq \tilde{M}$. By Corollary 3.7, for the CGN to have a better linear estimation, we need to ensure that $\tilde{M} < L_1$. This inequality can be rearranged to (12). $\square$

Proposition 3.9. *When using SDFEM discretization, and parameter ρ satisfies (12), then the linear estimation of CGN is better for every iterative step.*

Proof. This proof is analogous to the proof of Proposition 3.8. Let us consider the bilinear form $a(u, v)$ corresponding to the SDFEM:

$$a(u, v) := \int_{\Omega} (\varepsilon \nabla u \cdot \nabla v + (\rho \mathbf{w}_0 \cdot \nabla u) v + \delta (\rho \mathbf{w}_0 \cdot \nabla u) (\rho \mathbf{w}_0 \cdot \nabla v)).$$

We are looking for its upper bound with respect to the S -norm:

$$\langle u, v \rangle_S = \int_{\Omega} (\varepsilon \nabla u \cdot \nabla v + \delta (\rho \mathbf{w}_0 \cdot \nabla u) (\rho \mathbf{w}_0 \cdot \nabla v)).$$

The bilinear form is bounded as follows:

$$|a(u, v)| \leq |\langle u, v \rangle_S| + \left| \int_{\Omega} (\rho \mathbf{w}_0 \cdot \nabla u) v \right| \leq \left(1 + \frac{\rho \|\mathbf{w}_0\|_{L^\infty}}{\varepsilon \sqrt{2}\pi} \right) \|u\|_S \|v\|_S, \quad \forall u, v \in H_0^1(\Omega)$$

The same calculation can be repeated as in the previous proof. $\square$

Remark 3.10.

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

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

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

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

Proposition 3.11. *When using SDFEM discretization, and parameter ρ satisfies*

$$\rho > \frac{1}{L_1 - 1} \cdot \frac{C_{\mathbf{w}_0}}{\delta} \approx 0.574 \cdot \frac{C_{\mathbf{w}_0}}{\delta}, \quad (14)$$

where $C_{\mathbf{w}_0}$ is the constant in the streamline Poincaré–Friedrichs inequality, then the linear estimation of CGN is better for every iterative step.

Proof. We introduce a modified bilinear form $\tilde{a}(u, v)$ and inner product $\langle u, v \rangle_{\tilde{S}}$ as follows:

$$\begin{aligned} \tilde{a}(u, v) &:= \frac{1}{\rho} a(u, v) = \int_{\Omega} \left(\frac{\varepsilon}{\rho} \nabla u \cdot \nabla v + (\mathbf{w}_0 \cdot \nabla u) v + \rho \delta (\mathbf{w}_0 \cdot \nabla u) (\mathbf{w}_0 \cdot \nabla v) \right) \\ \langle u, v \rangle_{\tilde{S}} &:= \frac{1}{\rho} \langle u, v \rangle_S = \int_{\Omega} \left(\frac{\varepsilon}{\rho} \nabla u \cdot \nabla v + \rho \delta (\mathbf{w}_0 \cdot \nabla u) (\mathbf{w}_0 \cdot \nabla v) \right) \end{aligned}$$

These coincide with the weak form of a nonsymmetric elliptic problem with $\tilde{\varepsilon} := \frac{\varepsilon}{\rho}$, $\mathbf{w} := \mathbf{w}_0$ and parameter $\tilde{\delta} := \rho \delta$. We apply the streamline Poincaré–Friedrichs inequality (see [3], sec. 3.2.):

$$|\tilde{a}(u, v)| \leq \left(1 + \frac{C_{\mathbf{w}_0}}{\tilde{\delta}} \right) \|u\|_{\tilde{S}} \|v\|_{\tilde{S}}, \quad \forall u, v \in H_0^1(\Omega)$$

If we multiply both sides by ρ and substitute $\tilde{\delta}$, we get the desired inequality:

$$|a(u, v)| \leq \left(1 + \frac{C_{\mathbf{w}_0}}{\rho \delta} \right) \|u\|_S \|v\|_S, \quad \forall u, v \in H_0^1(\Omega)$$

This yields the upper bound

$$\hat{M} := 1 + \frac{C_{\mathbf{w}_0}}{\rho \delta}. \quad (15)$$

By equivalent preconditioning, $\|S^{-1}A\|_S$ inherits the upper bound of $a(u, v)$ owing to (4). That is, $\|S^{-1}A\|_S \leq \hat{M}$. By Corollary 3.7, for the CGN to have a better linear estimation, we need to ensure that $\hat{M} < L_1$. This inequality can be rearranged to (14). $\square$

Remark 3.12. 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, $\|S^{-1}A\|_S \rightarrow 1$ as $\rho \rightarrow \infty$.

Corollary 3.13. *When using SDFEM discretization, if*

$$\rho < \sqrt{2\pi}(L_1 - 1) \frac{\varepsilon}{\|\mathbf{w}_0\|_{L^\infty}} \quad \text{or} \quad \rho > \frac{C_{\mathbf{w}_0}}{(L_1 - 1)\delta}, \quad (16)$$

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

Proof. We combine the results of Proposition 3.9 and 3.11. □

Proposition 3.14. *When using SDFEM discretization with a parameter δ for which*

$$\delta > \frac{C_{\mathbf{w}_0} \|\mathbf{w}_0\|_{L^\infty}}{\sqrt{2}\pi(L_1 - 1)^2 \varepsilon} \approx 0.0741 \cdot \frac{C_{\mathbf{w}_0} \|\mathbf{w}_0\|_{L^\infty}}{\varepsilon}, \quad (17)$$

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

Proof. This is a consequence of Corollary 3.13. For any $\rho > 0$, at least one of the conditions in (16) is satisfied if

$$\frac{\sqrt{2}\pi(L_1 - 1)\varepsilon}{\|\mathbf{w}_0\|_{L^\infty}} > \frac{C_{\mathbf{w}_0}}{(L_1 - 1)\delta}.$$

This can be rearranged to (17). □

3.3 Application of the linear results to interpret the numerical experiments

Proposition 3.15. *For any nonsymmetric elliptic problem of the form (11), if $\rho = 0$, then both iterative methods converge after one step.*

Proof. If $\rho = 0$, then A is symmetric and $S = A$, therefore in the algorithms $r_1 = r_0 - r_0 = \mathbf{0}$, and $\|r_1\|_S = 0 < \text{TOL}$ is satisfied. □

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 3.10 we showed that $\|S^{-1}A\|_S \rightarrow 1$ as $\rho \rightarrow 0$, 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 3.8 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 3.7. 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 3.9 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 3.12 we showed that $\|S^{-1}A\|_S \rightarrow 1$ as $\rho \rightarrow \infty$. Therefore, we can use the previous argument with Proposition 3.11. $\blacksquare$

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 3.14. $\blacksquare$

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:

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

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

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

According to the relation in (12), 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 $\varepsilon = 1$.

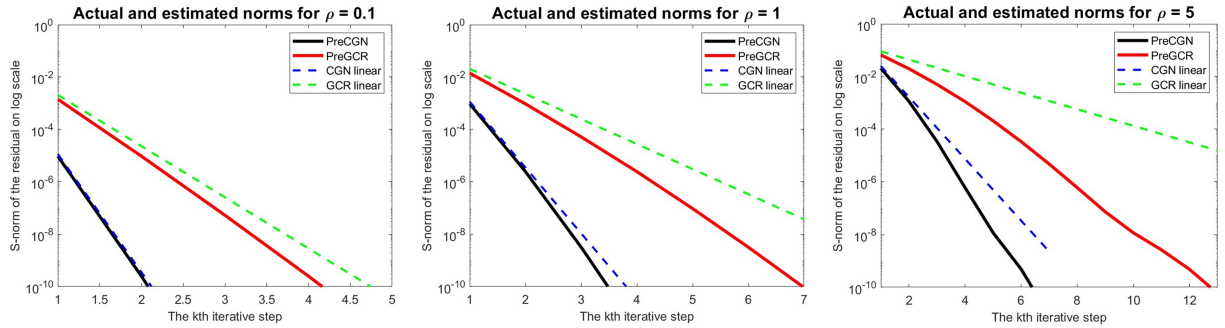


Figure 6: The actual and linear estimations of $\|r_k\|_S$ with FEM, $n = 20$, $\text{TOL} = 10^{-10}$, $\varepsilon = 1$, $\mathbf{w}_0 = (1, 0)$ and $\rho = 0.1, 1, 5$.

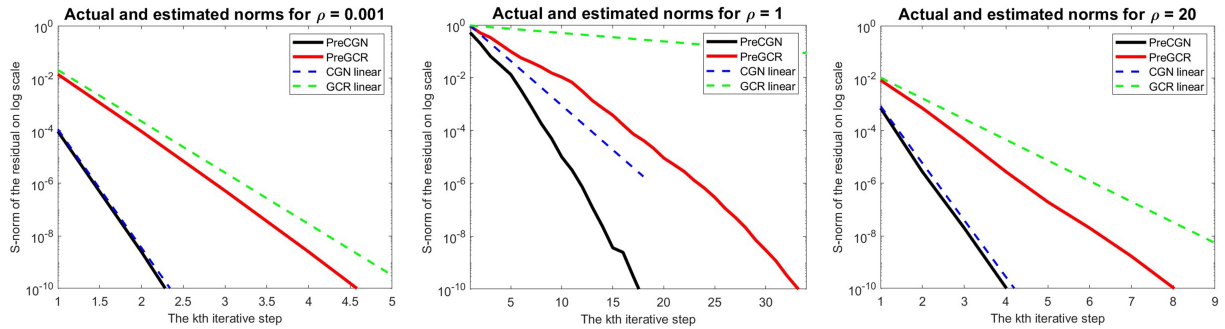


Figure 7: The actual and linear estimations of $\|r_k\|_S$ with SDFEM, $n = 20$, $\text{TOL} = 10^{-10}$, $\delta = h$, $\varepsilon = 0.01$, $\mathbf{w}_0 = (1, 0)$ and $\rho = 0.001, 1, 20$.

According to the relation in (16), the CGN method is better in the domain $\mathbb{R}^+ \setminus [7.741 \cdot \varepsilon, \frac{0.812}{\delta}]$ when the discretization is executed with the SDFEM, which is $\mathbb{R}^+ \setminus [0.0774, 16.917]$ for $\varepsilon = 0.01$ and $\delta = 0.048$.

In order to plot the linear estimations, we need to calculate the actual value of $\|S^{-1}A\|_S$ for any parameters. The following proposition helps to calculate it using the Euclidean norm, which is built in MATLAB.

Proposition 3.16. *The S -norm of $S^{-1}A$ can be calculated as $\|S^{-1}A\|_S = \|S^{-\frac{1}{2}}AS^{-\frac{1}{2}}\|$.*

Proof. Let $S^{-\frac{1}{2}}$ be the principal square root of S^{-1} and $d := S^{\frac{1}{2}}c$ for an arbitrarily fixed $c \in \mathbb{R}^N$.

$$\begin{aligned}\|S^{-1}Ac\|_S^2 &= \langle Ac, S^{-1}Ac \rangle = \left\langle AS^{-\frac{1}{2}}d, S^{-\frac{1}{2}} \left(S^{-\frac{1}{2}}AS^{-\frac{1}{2}}d \right) \right\rangle = \|S^{-\frac{1}{2}}AS^{-\frac{1}{2}}d\|^2; \\ \|c\|_S^2 &= \langle Sc, c \rangle = \left\langle S^{\frac{1}{2}} \left(S^{\frac{1}{2}}c \right), c \right\rangle = \|S^{\frac{1}{2}}c\|^2 = \|d\|^2; \\ \|S^{-1}A\|_S &= \sup_{c \neq 0} \frac{\|S^{-1}Ac\|_S}{\|c\|_S} = \sup_{d \neq 0} \frac{\|S^{-\frac{1}{2}}AS^{-\frac{1}{2}}d\|}{\|d\|} = \|S^{-\frac{1}{2}}AS^{-\frac{1}{2}}\|. \quad \square\end{aligned}$$

For large matrices, it is slow to calculate their norm, so for optimizational purposes we use grid density $n = 20$ and $\delta = h \approx 0.048$. The plot of the residual norms $\|r_k\|_S$ and the corresponding linear estimations can be seen in figure 6 for $\varepsilon = 1$ with FEM and in figure 7 for $\varepsilon = 0.01$ with SDFEM. These graphs confirm that according to our calculations if $\rho \in (0, 7.741)$ in figure 6 and $\rho \in \mathbb{R}^+ \setminus [0.0774, 16.917]$ in figure 7, 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 6 start deteriorating as ρ is increased, and in case of SDFEM in figure 7, the linear estimations do not approximate well when $\rho = 1$, which is in the excluded interval.

3.4 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 in (7) and (8).

Proposition 3.17. $\|A^{-1}S\|_S \leq 1$.

Proof. In Proposition 3.6 we have verified that $\|c\|_S^2 = \langle Ac, c \rangle$. We can use the Cauchy–Schwarz inequality:

$$\begin{aligned}\|c\|_S^2 &= \langle S^{-1}Ac, c \rangle_S \leq \|S^{-1}Ac\|_S \|c\|_S \\ \Rightarrow \|c\|_S &\leq \|S^{-1}Ac\|_S\end{aligned}$$

Let us introduce a new variable $v := S^{-1}Ac$. Then $c = (S^{-1}A)^{-1}v = A^{-1}Sv$. If we rewrite the inequality, we get $\|A^{-1}Sv\|_S \leq \|v\|_S$ implying that $\|A^{-1}S\|_S \leq 1$. $\square$

Remark 3.18. In the numerical tests, $\|A^{-1}S\|_S = 1$ was satisfied as well.

Proposition 3.19. *Let $\|A^{-1}S\|_S = 1$ be satisfied, and suppose that E is antisymmetric. The superlinear estimation of GCR in the k th iterative step is better than that of CGN, that is, the upper bound for the convergence rate in (8) is lower than the one in (9) exactly when*

$$\sum_{j=1}^k s_j(E) < 2 \sum_{j=1}^k s_j^2(E). \quad (18)$$

Proof. We need to compare the superlinear estimations in (8) and (9):

$$\frac{\|A^{-1}S\|_S}{k} \sum_{j=1}^k s_j(E) < \frac{2\|A^{-1}S\|_S^2}{k} \sum_{j=1}^k s_j^2(E)$$

If we use $\|A^{-1}S\|_S = 1$ and multiply both sides by k , then we obtain (18). $\square$

In order to use Proposition 3.19 similar to the linear estimations, we need to give $s_j(E)$ in terms of the studied parameters. Remark 3.22 shows that this is not possible in general. What we can do instead is that we calculate the singular values when $\varepsilon = 0$, which provides a reasonable approximation for convection-dominated problems with ε close to zero.

Definition 3.20. When using SDFEM discretization, let operator $A : H_0^1(\Omega) \rightarrow H_0^1(\Omega)$ be defined as

$$\langle Au, v \rangle_S = \int_{\Omega} \left(\varepsilon \nabla u \cdot \overline{\nabla v} + (\rho \mathbf{w}_0 \cdot \nabla u) \bar{v} + \delta (\rho \mathbf{w}_0 \cdot \nabla u) (\rho \mathbf{w}_0 \cdot \overline{\nabla v}) \right) \quad (\forall u, v \in H_0^1(\Omega)),$$

and operator E be defined by the decomposition $A = I + E$, where then

$$\langle Eu, v \rangle_S = \int_{\Omega} (\rho \mathbf{w}_0 \cdot \nabla u) \bar{v} \quad (\forall u, v \in H_0^1(\Omega)).$$

Proposition 3.21. *When using SDFEM discretization, if $\varepsilon = 0$ and $\mathbf{w}_0 = (1, 0)$, then the eigenvalues of operator E are*

$$\mu_j(E) = \frac{i}{2\pi\rho\delta j} \quad (j \in \mathbb{Z} \setminus \{0\}). \quad (19)$$

Proof. This proof is analogous to Example 1 in [7]. In this special case, for any $u, v \in H_0^1(\Omega)$:

$$\begin{aligned} \langle u, v \rangle_S &= \int_{\Omega} \left(\varepsilon \nabla u \cdot \overline{\nabla v} + \delta (\rho \mathbf{w}_0 \cdot \nabla u) (\rho \mathbf{w}_0 \cdot \overline{\nabla v}) \right) = \delta \rho^2 \int_{\Omega} \partial_x u \overline{\partial_x v} = -\delta \rho^2 \int_{\Omega} \partial_x^2 u \bar{v} \\ \langle Eu, v \rangle_S &= \int_{\Omega} (\rho \mathbf{w}_0 \cdot \nabla u) \bar{v} = \rho \int_{\Omega} \partial_x u \bar{v} \end{aligned}$$

Hence, the weak form of the related eigenvalue problem is $\langle Eu, v \rangle_S = \langle \mu u, v \rangle_S$. The strong form of this problem is

$$\begin{cases} \partial_x u = -\mu \delta \rho \partial_x^2 u \\ u|_{\partial\Omega} = 0 \end{cases}$$

The solutions of the PDE are in the form

$$u(x, y) = -v(y) \mu \delta \rho e^{-\frac{1}{\mu \delta \rho} x} + z(y).$$

We need to enforce the boundary conditions:

$$\begin{aligned} 0 = u(0, y) = -v(y)\mu\delta\rho + z(y) &\Rightarrow u(x, y) = z(y) \left(1 - e^{-\frac{1}{\mu\delta\rho}x}\right) \\ 0 = u(1, y) = z(y) \left(1 - e^{-\frac{1}{\mu\delta\rho}}\right), \quad \forall z \in H_0^1(0, 1) \end{aligned}$$

Consequently, the eigenvalues are determined by

$$1 - e^{-\frac{1}{\mu\delta\rho}} = 0 \Leftrightarrow -\frac{1}{\mu_j\delta\rho} = 2\pi ji \Leftrightarrow \mu_j = \frac{i}{2\pi\rho\delta j} \quad (j \in \mathbb{Z} \setminus \{0\}) \quad \square$$

Remark 3.22. If $\varepsilon > 0$ and $\mathbf{w}_0 = (1, 0)$, then the strong form of the eigenvalue problem is

$$\begin{cases} \rho \partial_x u = -\mu(\varepsilon + \delta\rho^2) \partial_x^2 u - \mu\varepsilon \partial_y^2 u \\ u|_{\partial\Omega} = 0 \end{cases},$$

which cannot be solved analytically in general.

Proposition 3.23. When using SDFEM discretization, if $\varepsilon = 0$ and $\mathbf{w}_0 = (1, 0)$, then

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

Proof. Operator E is antisymmetric, that is

$$\langle Eu, v \rangle_S = \int_{\Omega} (\rho \mathbf{w}_0 \cdot \nabla u) \bar{v} = - \int_{\Omega} u (\rho \mathbf{w}_0 \cdot \nabla \bar{v}) = - \langle u, Ev \rangle_S \quad (\forall u, v \in H_0^1(\Omega)).$$

Therefore, the singular values of E are the absolute values of the eigenvalues. Every singular value has a multiplicity of two because $|\mu_j| = |\mu_{-j}|$. Hence, the singular values in decreasing order:

$$s_{2j}(E) = s_{2j-1}(E) = |\mu_j| = \left| \frac{i}{2\pi\rho\delta j} \right| = \frac{1}{2\pi\rho\delta j} \quad (j = 1, 2, \dots) \quad \square$$

Proposition 3.24. 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

$$\rho < \frac{1}{\pi\delta} \frac{\sum_{j=1}^k \frac{1}{j^2}}{\sum_{j=1}^k \frac{1}{j}} \approx \frac{1}{\pi\delta} \frac{\frac{\pi^2}{6} - \frac{1}{k}}{0.5772 + \ln k + \frac{1}{2k}}. \quad (20)$$

Proof. We use Proposition 3.19 with the singular values calculated in Proposition 3.23.

$$\sum_{j=1}^{k'} s_j(E) = 2 \sum_{j=1}^k s_{2j}(E) = 2 \sum_{j=1}^k \frac{1}{2\pi\rho\delta j} = \frac{1}{\pi\rho\delta} \sum_{j=1}^k \frac{1}{j} \quad (21)$$

$$2 \sum_{j=1}^{k'} s_j^2(E) = 4 \sum_{j=1}^k s_{2j}^2(E) = 4 \sum_{j=1}^k \frac{1}{4(\pi\rho\delta)^2 j^2} = \frac{1}{(\pi\rho\delta)^2} \sum_{j=1}^k \frac{1}{j^2} \quad (22)$$

We need to compare the two quantities:

$$\sum_{j=1}^{k'} s_j(E) < 2 \sum_{j=1}^{k'} s_j^2(E) \Leftrightarrow \rho \sum_{j=1}^k \frac{1}{j} < \frac{1}{\pi\delta} \sum_{j=1}^k \frac{1}{j^2}$$

This can be rearranged to the form in (20). The partial sums are approximated by the Euler–Maclaurin formula with a second-order error. $\square$

Remark 3.25. The ratio of the sums in (20) is strictly monotonic decreasing, which gives a decreasing upper bound for ρ in Proposition 3.24 as k is increasing. The ratio of the sums and its approximation can be seen in figure 8. Proposition 3.24 is also in accordance with the middle part of the graphs in figure 5 because a small value of ρ gives advantage to the GCR method. It also does not contradict with the first part of the graphs because the linear estimations are more relevant in that section.

Definition 3.26. When using FEM discretization, let operator $A : H_0^1(\Omega) \rightarrow H_0^1(\Omega)$ be defined as

$$\langle Au, v \rangle_S = \int_{\Omega} \left(\varepsilon \nabla u \cdot \overline{\nabla v} + (\rho \mathbf{w}_0 \cdot \nabla u) \bar{v} \right) \quad (\forall u, v \in H_0^1(\Omega)),$$

and operator E and E_0 be defined by the decomposition $A = I + E = I + \rho E_0$, where then

$$\langle Eu, v \rangle_S = \int_{\Omega} (\rho \mathbf{w}_0 \cdot \nabla u) \bar{v} = \rho \int_{\Omega} (\mathbf{w}_0 \cdot \nabla u) \bar{v} = \rho \langle E_0 u, v \rangle_S \quad (\forall u, v \in H_0^1(\Omega)).$$

Proposition 3.27. *Let $k \in \mathbb{N}$ be an arbitrary index. When using FEM discretization, the superlinear estimation of the GCR method in the k th iterative step is better than that of the CGN method if*

$$\rho > \frac{\sum_{j=1}^k s_j(E_0)}{2 \sum_{j=1}^k s_j^2(E_0)}. \quad (23)$$

Proof. We use Proposition 3.19 and the relation $s_j(E) = s_j(\rho E_0) = \rho s_j(E_0)$:

$$\sum_{j=1}^k s_j(E) < 2 \sum_{j=1}^k s_j^2(E) \quad \Leftrightarrow \quad \sum_{j=1}^k s_j(E_0) < 2\rho \sum_{j=1}^k s_j^2(E_0) \quad \square$$

Remark 3.28. Proposition 3.27 supports Observation 2 because as ρ is increased, (23) is satisfied for more and more indices, which gives advantage to the GCR method. However, as pointed out in section 3.5 with figure 9, in most cases the superlinear estimations do not approximate well the residual norms when ρ is large and standard FEM discretization is used.

3.5 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 3 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 4 and 5 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 Remark 3.28 in case of FEM and in Observation 5 and Remark 3.25 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 9 and 10.

When $\varepsilon = 0$, the previous calculations show why the superlinear estimation deteriorates as ρ approaches zero. In equation (21) we can see that by dividing ρ , the sum of the singular values becomes very large if ρ is very small. In equation (22) we divide by ρ^2 , which makes this sum even larger. This can be seen also in figure 10.

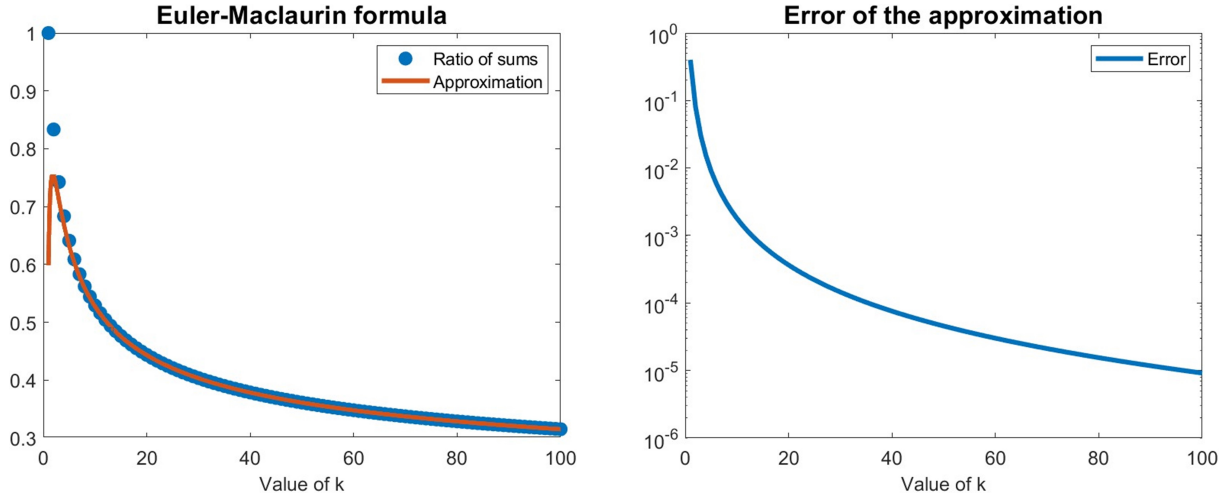


Figure 8: The ratio of sums in (20) and its approximation with the Euler–Maclaurin formula.

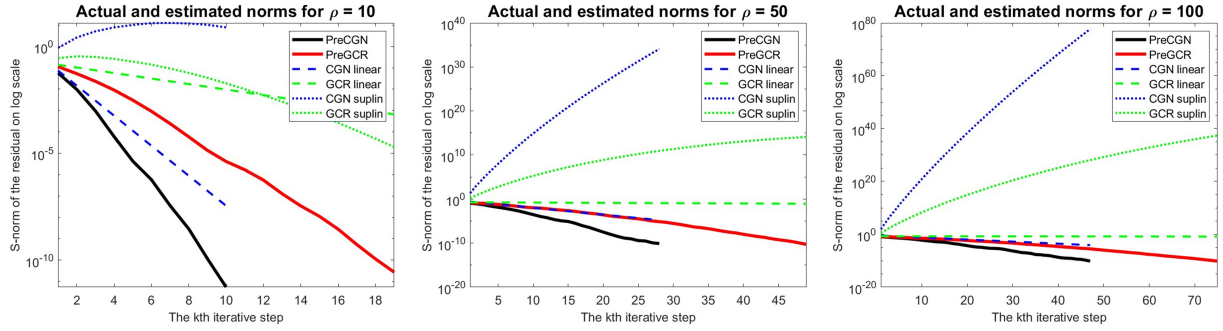


Figure 9: The actual residual norms and their linear and superlinear estimations with FEM, $n = 20$, $\text{TOL} = 10^{-10}$, $\varepsilon = 1$, $\mathbf{w}_0 = (1, 0)$ and $\rho = 10, 50, 100$.

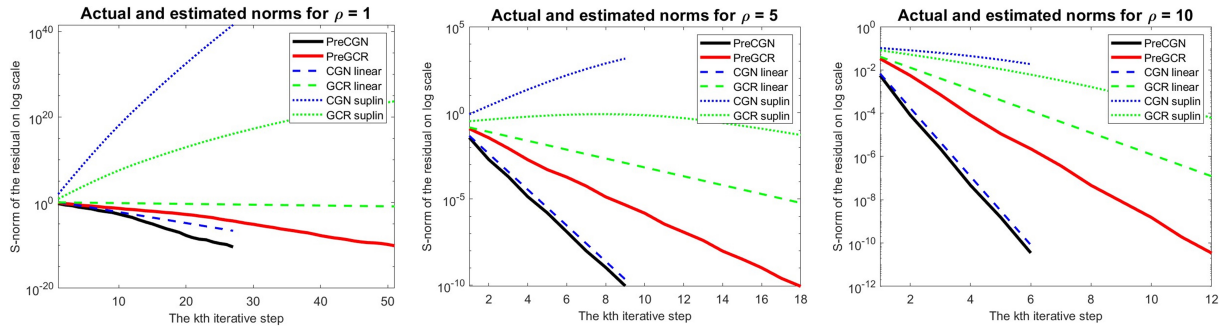


Figure 10: The actual residual norms and their linear and superlinear estimations with SDFEM, $n = 20$, $\text{TOL} = 10^{-10}$, $\delta = h$, $\varepsilon = 0$, $\mathbf{w}_0 = (1, 0)$ and $\rho = 1, 5, 10$.

Finally, we show two extreme cases where the stronger superlinear estimation of the GCR in (10) is usable according to the numerical tests. One way to find such cases is to lower the tolerance to $\text{TOL} = 10^{-100}$. This raises problems with the implementation of the iterative methods because MATLAB uses only 16 digits of precision by default, which makes the GCR method unable to decrease $\|r_k\|_S$ even below 10^{-10} in most cases. This problem is examined in [8], where it is suggested that a higher arithmetic precision could restore the convergence. This can be achieved by using the `vpa()` function in the Symbolic Math Toolbox to set the precision to 100 digits.

In figure 11, we can see the residual norms for $\rho = 1$ with SDFEM discretization, where the linear estimations are not good approximators as we have seen before. The grid parameter is set to $n = 10$, which enables us to see the full convergence of the iterative methods because they must find the exact solution in at most $n^2 = 100$ steps in case of GCR and in at most $\frac{n^2}{2} = 50$ steps in case of CGN. We can see that on this larger scale, the superlinear estimation of GCR is a reasonable approximation of the residual norm. This can be seen also in figure 12 for $\rho = 20$ with FEM discretization.

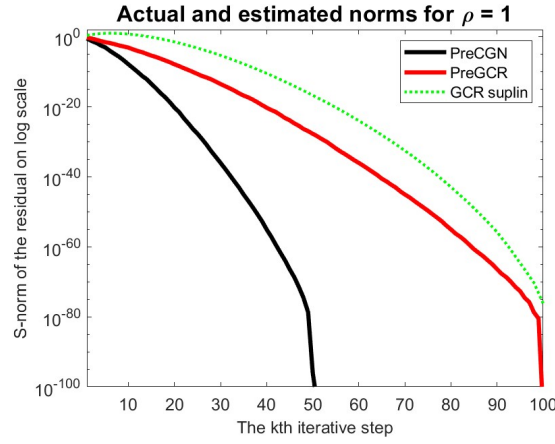


Figure 11: The actual residual norms and the superlinear estimation of GCR with (10) for SDFEM, $n = 10$, $\text{TOL} = 10^{-100}$, $\delta = h$, $\varepsilon = 0.01$, $\mathbf{w}_0 = (1, 0)$ and $\rho = 1$.

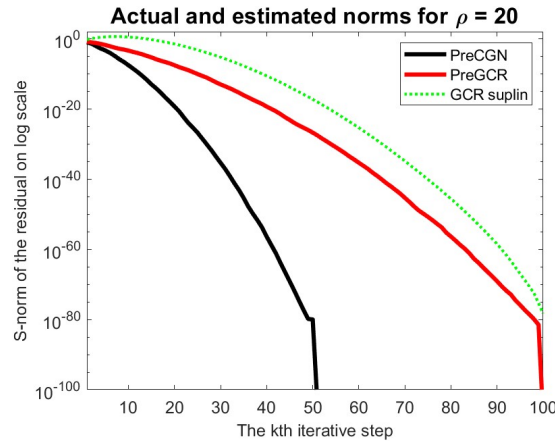


Figure 12: The actual residual norms and the superlinear estimation of GCR with (10) for FEM, $n = 10$, $\text{TOL} = 10^{-100}$, $\varepsilon = 1$, $\mathbf{w}_0 = (1, 0)$ and $\rho = 20$.

3.6 Conclusions

In this thesis we have compared the performance of the CGN and GCR methods when they are applied to the discretization of stationary convection-diffusion type nonsymmetric elliptic problems, and the resulting system of equations is preconditioned with the symmetric part of the matrix. We showed by numerical experiments that either method could prevail for different coefficients of the PDE when using any of the three studied discretization methods.

In order to explain the outcomes, we compared the linear and superlinear estimations of the residual norm. 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 numerical tests confirmed that in these intervals the linear estimations are reasonable approximations of the actual residual norms, therefore the theoretical results explain why the CGN method converges faster in these cases.

The superlinear estimations were compared for a constant vector field in the limiting case when the diffusion parameter is zero, and an upper bound was obtained for the parameter of the convection term in case of SDFEM discretization. The thesis was concluded by numerical experiments showing the limitations of using the superlinear estimations as an approximation of the residual norms, and we provided some special cases when a stronger superlinear estimation of the GCR could be used. Our further goal is to find a way to explain the remaining cases using the superlinear estimations.

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] KARÁTSÓN, J.; HORVÁTH, R.: *Numerical Methods for Elliptic Partial Differential Equations*. URL: <https://kajkaat.web.elte.hu/pdnmell-ang-2022.pdf>
- [3] AXELSSON, O.; KARÁTSÓN, 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.; KARÁTSÓN, 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] KARÁTSÓN, 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/>