Comparison of iterative methods for discretized nonsymmetric elliptic problems

Math Project III.

Lados Bálint István

Supervisor: **Karátson János**

Eötvös Loránd University

Department of Applied Analysis and Computational Mathematics

9 January, 2025

The studied elliptic problem

Let us consider the following elliptic boundary value problem:

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

where $\Omega = (0,1)^2$ is the unit square, $\varepsilon > 0$ is a constant, $\mathbf{w} \in C^1(\overline{\Omega}, \mathbb{R}^2)$ is a divergence-free vector field and $f \in L^2(\Omega)$.

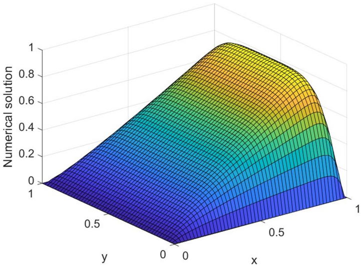
- This models a stationary convection-diffusion process.
- It is related to the linearized version of the Navier–Stokes equations arising from fluid dynamics.
- Convection-dominated problems form an important subclass: $\varepsilon \ll 1$.
- The problem has a unique weak solution $u \in H_0^1(\Omega)$ such that

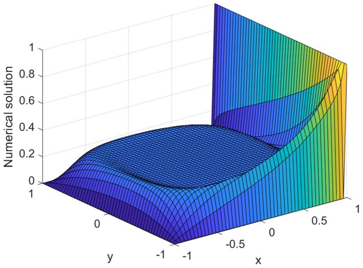
  $\int_\Omega (\varepsilon\nabla u\cdot\nabla v + (\mathbf{w}\cdot\nabla u)\,v) = \int_\Omega fv \qquad (\forall v \in H_0^1(\Omega)).$

Discretization methods

We approximate the solution with one of the following numerical methods on the uniform grid of the unit square:

- **FDM:** Finite difference method, second-order central scheme.
- **FEM:** Finite element method, first-order Courant elements.
- **SDFEM:** Streamline diffusion finite element method for convection-dominated problems with a stabilizing parameter $\delta > 0$.





Nonsymmetric iterative methods

The discretization leads to a system of linear equations $Au = b$, where matrix A is nonsymmetric.

This can be solved by one of the following iterative methods:

- **CGN:** The conjugate gradient method applied to the normal equation.
- **GCR:** Minimization of the residual error in the Krylov subspace.

Preconditioning: In order to boost the rate of convergence, we solve $S^{-1}Au = S^{-1}b$ instead of the original system of equations, where $S := \frac{A+A^T}{2}$ is the symmetric part of matrix A.

Stop criterion: $\|r_n\|_S := \sqrt{\langle Sr_n, r_n\rangle} < \mathrm{TOL}$, i.e. when the S-norm of the residual error vector decreases below a given threshold (e.g. $\mathrm{TOL} = 10^{-10}$).

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**

$z_n := S^{-1}Ap_n$;

$\alpha_n = -\dfrac{\|s_n\|_S^2}{\|z_n\|_S^2}$;

$u_{n+1} := u_n + \alpha_n p_n$;

$r_{n+1} := r_n + \alpha_n z_n$;

$s_{n+1} := S^{-1}A^T r_{n+1}$;

$\beta_n = \dfrac{\|s_{n+1}\|_S^2}{\|s_n\|_S^2}$;

$p_{n+1} := s_{n+1} + \beta_n p_n$;

$n := n+1$;

end

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**

$\alpha_n := \dfrac{\langle r_n, z_n\rangle_S}{\|z_n\|_S^2}$;

$u_{n+1} := u_n + \alpha_n p_n$;

$r_{n+1} := r_n - \alpha_n z_n$;

$s_n := S^{-1}Ar_{n+1}$;

for $i = 0, 1, \ldots, n$ **do**

$\beta_{i,n} := -\dfrac{\langle s_n, z_i\rangle_S}{\|z_i\|_S^2}$;

end

$p_{n+1} := r_{n+1} + \sum_{i=0}^{n} \beta_{i,n} p_i$;

$z_{n+1} := s_n + \sum_{i=0}^{n} \beta_{i,n} z_n$;

$n := n+1$;

end

Comparison of the two iterative methods

Let us consider the following problem depending on the parameters $\varepsilon > 0$ and $\rho > 0$:

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

Question: Which iterative method solves the resulting system of linear equations $Au = b$ in less iterative steps for different values of ε and ρ?

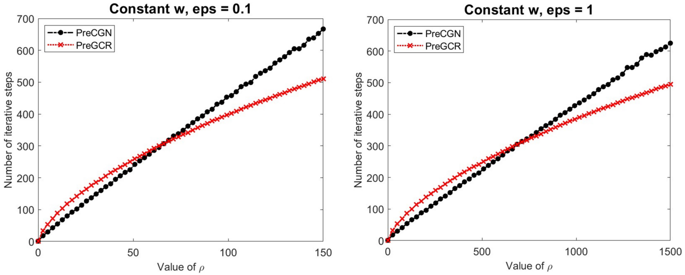
Test problem: Consider the constant vector field $\mathbf{w}_0 := (1, 0)$.

Numerical test: We fix the value of ε, increase ρ starting from 0, and plot the number of iterative steps until convergence for the three discretization methods separately.

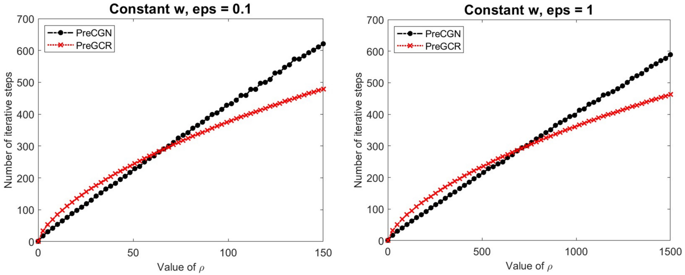
Implementation: MATLAB

Numerical results: FDM and FEM

FDM:



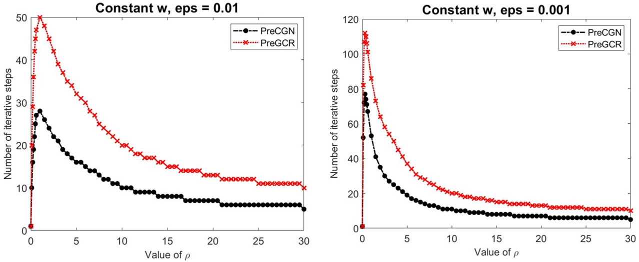
FEM:



Numerical results: SDFEM

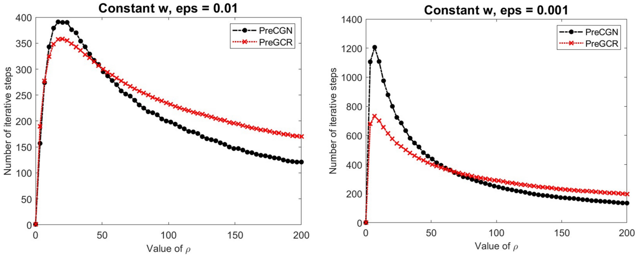
SDFEM

δ = 10^{-2}



SDFEM

δ = $5*10^{-5}$



Linear convergence estimates

In order to explain the numerical results, I used the following two well-known linear convergence estimates:

$$\textbf{CGN:} \qquad \left(\frac{\|r_k\|_S}{\|r_0\|_S}\right)^{\frac{1}{k}} \leq 2^{\frac{1}{k}} \, \frac{M-m}{M+m} \qquad (k = 1, \ldots, N)$$

$$\textbf{GCR:} \qquad \left(\frac{\|r_k\|_S}{\|r_0\|_S}\right)^{\frac{1}{k}} \leq \sqrt{1 - \left(\frac{m}{M}\right)^2} \qquad (k = 1, \ldots, N)$$

Here, $r_k = S^{-1}Au_k - S^{-1}b$ is the residual error in the kth iterative step, and $M \geq m > 0$ are constants defined in the following way:

$$m := \inf\{\langle Ac, c\rangle : \ \|c\|_S = 1\}$$

$$M := \|S^{-1}A\|_S = \sup\{\langle Ac, d\rangle : \ \|c\|_S = \|d\|_S = 1\}$$

A general theoretical result for the linear estimates

Theorem: Let $k \in \{1, \ldots, N\}$ be an arbitrary index. The linear estimation of the GCR method in the kth iterative step is better than that of the CGN method if and only if $\frac{M}{m} > L_k$, where 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

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

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
L_k	2.7423	5.5708	8.4388	11.3158	14.1962	17.0783	19.9614

Consequences of the theorem

Corollary 1: In case of standard FEM discretization, if

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

then the linear estimation of the CGN method is better in each step.

Corollary 2: In case of SDFEM discretization, if

$$\rho < 7.741 \cdot \frac{\varepsilon}{\|\mathbf{w}_0\|_{L^\infty}} \quad \text{or} \quad \rho > 0.574 \cdot \frac{C_{\mathbf{w_0}}}{\delta},$$

then the linear estimation of the CGN method is better in each step.

Corollary 3: In case of SDFEM discretization, if

$$\delta > \frac{C_{\mathbf{w_0}}\|\mathbf{w}_0\|_{L^\infty}}{\sqrt{2}\pi(L_1 - 1)^2\varepsilon} \approx 0.0741 \cdot \frac{C_{\mathbf{w_0}}\|\mathbf{w}_0\|_{L^\infty}}{\varepsilon},$$

then the linear estimation of the CGN method is better in each step for any $\rho > 0$.

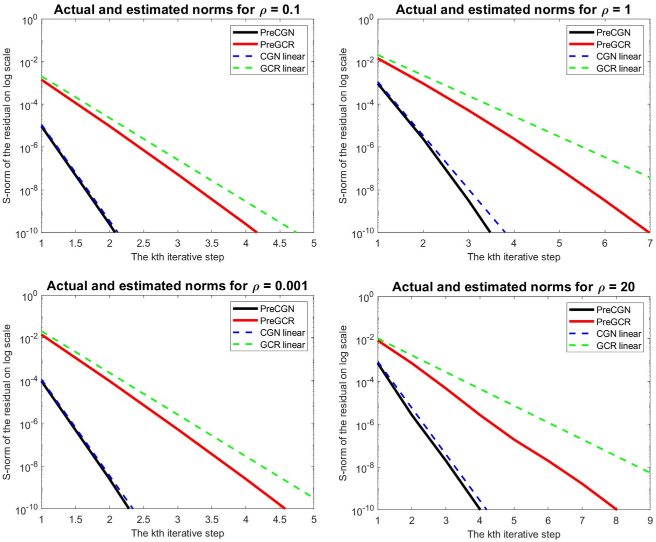
Example: In case of $\mathbf{w}_0 = (1,\, 0)$, $\|\mathbf{w}_0\|_{L^\infty} = 1$ and $C_{\mathbf{w_0}} = \sqrt{2}$.

The actual residual norms and their linear estimation

FEM

SDFEM

$\delta = 10^{-2}$



Superlinear convergence estimates

The linear estimates could not explain those parts of the graphs where the GCR method performs better than the CGN method.

For further examination, I used the following two well-known superlinear convergence estimates:

$$\textbf{CGN:} \quad \left(\frac{\|r_k\|_S}{\|r_0\|_S}\right)^{\frac{1}{k}} \le \frac{2\|A^{-1}S\|_S^2}{k} \sum_{j=1}^{k} s_j^2(E) \qquad (k = 1, \dots, N)$$

$$\textbf{GCR:} \quad \left(\frac{\|r_k\|_S}{\|r_0\|_S}\right)^{\frac{1}{k}} \le \frac{\|A^{-1}S\|_S}{k} \sum_{j=1}^{k} s_j(E) \qquad (k = 1, \dots, N)$$

Here, $S^{-1}A = I + E$, where E is an antisymmetric matrix, and $s_j(E)$ is the jth singular value of matrix E in decreasing order and with multiplicity.

Two specific results

Proposition: Let $k \in \{1, \ldots, N\}$ be an arbitrary index 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

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

Proposition: Let $k \in \{1, \ldots, N\}$ be an arbitrary index. When using standard FEM discretization, the superlinear estimation of the GCR method in the kth iterative step is better than that of the CGN method if

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

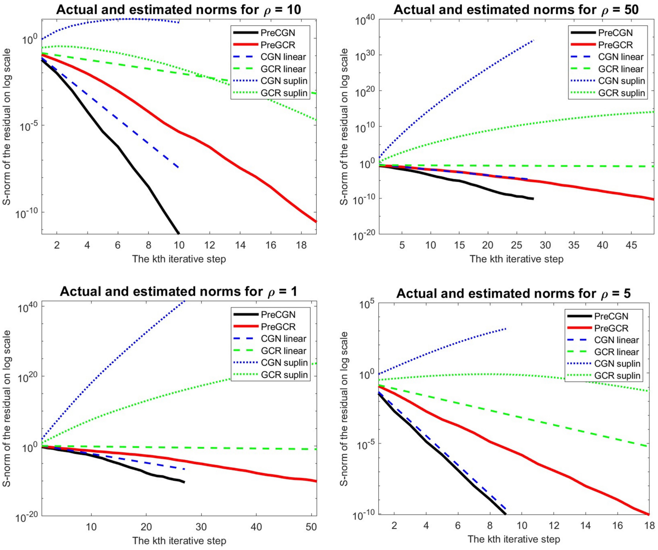
where $E_0 := \frac{1}{\rho} E$.

Problem with the superlinear estimations

FEM

SDFEM

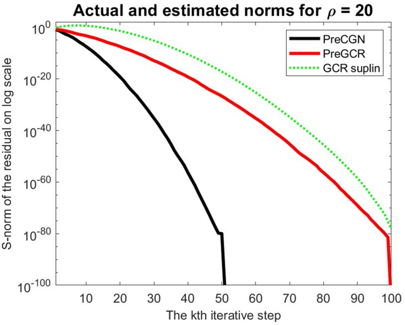
$\delta = 10^{-2}$

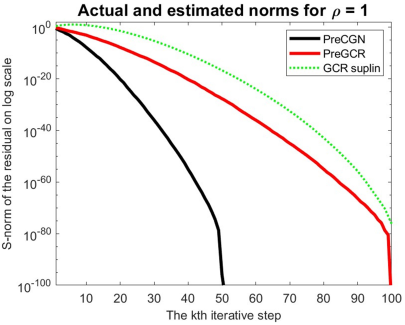


A possible improvement: stronger estimate, higher accuracy

There exists a stronger superlinear estimation of the GCR method:

$$\textbf{GCR:} \quad \left(\frac{\|r_k\|_S}{\|r_0\|_S}\right)^{\frac{1}{k}} \leq \|A^{-1}S\|_S \left(\prod_{j=1}^{k} s_j(E)\right)^{\frac{1}{k}} \qquad (k = 1, \ldots, N)$$





References

Thank you for your attention!

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.

Saad, Y.:
Iterative methods for sparse linear systems.
SIAM, Philadelphia, 2003.

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.

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