THEORETICAL AND NUMERICAL ANALYSIS OF A MINIMAL RESIDUAL SOLVER FOR 2D BOLTZMANN TRANSPORT EQUATION.

S. AKESBI AND E. MAITRE*

Abstract. Relying on the splitting of the collision operator introduced by [6] [1], we prove theoretical convergence for an infinite dimensional adaptation of the minimal residual algorithm for Boltzmann transport equation in dimension two. Then we compare this solver with known ones from a numerical point of view.

1. Introduction and notations. The behavior of neutrons in a two dimensional domain D, in interaction with them, is described by a function $f(x,\Omega)$ which represents, up to some factor, the flux of neutron density at the position x with velocity $\Omega \in B(0,1)$. A function $\sigma(x)$ accounts for neutron-domain interaction, whereas a kernel $k(x,\Omega,\Omega')$ describes collisions between neutrons. At last, a neutron source is represented by a non-negative function $S(x,\Omega)$. We refer to [10] and [19] for a more precise introduction.

The function f verifies an integro-differential equation. Our aim is to prove the convergence of a minimal residual method to solve this equation. Then we compare numerically our algorithm with some iterative methods developed in the last few years by S. Akesbi and M. Nicolet [5]. Note that the convergence of this algorithm could be accelerated by an adapted DSA [12].

1.1. Mathematical setting. Let D bet a bounded open set of $\mathbf{R}^2$ with lipschitz boundary ∂D, and $Q = D \times B$ where $B = B(0,1) = \left\{\Omega \in \mathbf{R}^2, \quad \|\Omega\|_2 < 1\right\}$. The outer normal $\mathbf{n}(\mathbf{x})$ to ∂D exists almost everywhere, and we define

$$\Gamma^- := \{(x,\Omega) \in \partial D \times B, \quad \Omega \cdot \mathbf{n}(x) < 0\}.$$

We consider the following problem : given a source term S, find $f : Q \to \mathbf{R}$ solution of the transport equation

$$(P) \quad \begin{cases} Tf(x,\Omega) = Kf(x,\Omega) + S(x,\Omega) & \text{in } Q, \\ f(x,\Omega) = 0 & \text{on } \Gamma^-, \end{cases}$$

where T is the transport operator, $Tf(x,\Omega) = \Omega \cdot \nabla_x f(x,\Omega) + \sigma(x) f(x,\Omega)$ whose domain is

$$\mathcal{D}(T) = \left\{ f \in L^2(Q) \ : \ \Omega \cdot \nabla_x f \in L^2(Q), \ f = 0 \text{ on } \Gamma^- \right\},$$

and K an integral operator of *positive* kernel k :

$$Kf(x,\Omega) = \int_B k(x,\Omega,\Omega') f(x,\Omega') d\Omega'.$$

We make the following

Assumptions :

(A1) $\sigma \in L^\infty(D)$, $\exists \sigma_0 > 0, \quad \sigma(x) \geq \sigma_0$ a.e. on D.

*Laboratoire de Mathématiques et Application, Université de Haute-Alsace, 4, rue des frères Lumière, F-68093 Mulhouse Cedex. e-mail : S.Akesbi@univ-mulhouse.fr, E.Maitre@univ-mulhouse.fr

(A2) $k(x,\Omega,\Omega') = k(x,\Omega',\Omega)$ and k is positive.

(A3) $\exists c \in [0,1), \quad \forall i \in \{1,2,3,4\}, \quad \int_{B_i} k(x,\Omega,\Omega')d\Omega' \leq \frac{\sigma_0 c}{4}$ a.e. on Q, where B_i is the i-th quarter of the disk B, see figure 1.

(A4) $k(x,\Omega,\Omega') = C(x)\sum_{l=1}^{N_k} a_l(\Omega)a_l(\Omega')$.

REMARK 1.

1. *One can replace assumption (A1) by this less restrictive assumption :*
(A1') $(f,g) \to \int_D \sigma(x)f(x)g(x)dx$ is a scalar product on $L^2(D)$.
This allows σ to vanish on sets of null measure in D. In this case one has to work in Lebesgue space with weight σ.

2. *Assumption (A4) is not used for theoretical proof of convergence. However, it is necessary to assume this form for k for the numerical splitting method to work (see numerical results). In this case the symmetry assumption of (A2) is automatically verified.*

3. *We can also replace (A3) by (A3)' : $\exists c \in [0,1), \quad k(x,\Omega,\Omega') \leq \frac{\sigma_0 c}{\pi}$*

4. *These assumptions (including (A4)) are satisfied for usual kernels of neutronic as the constant and Thomson kernels.*

5. *Assumptions (A1)-(A3) ensure the existence and uniqueness of the solution of (P) in $\mathcal{D}(T)$. Indeed, they are stronger than those of [10], for example the symmetry property of k with assumption (A3) give assumptions 2.67 p. 1105, which with (A1) imply assumption 2.40 p. 1092. From theorem 2 p. 1087 we know that $Af := \Omega\cdot\nabla_x f$ is a m-accretive operator with domain $\mathcal{D}(T)$, and the previous assumptions give existence and uniqueness of a solution of (P) in $\mathcal{D}(T)$ (theorem 4 p. 1105).*

6. *Note that all obtained results are valid for non zero incoming flux in (P). Note also that from the m-accretivity of A and assumption (A1), T^{-1} exists.*

1.2. Classical and splitting methods. The standard method to solve (P), called the source iteration method, is based on a decoupling between the differential and integral parts, through the following iterative scheme : given $f^0 \in \mathcal{D}(T)$, solve

$$(P_s) \quad \begin{cases} Tf^{n+1} = Kf^n + S & \text{in } Q, \\ f^{n+1} \in \mathcal{D}(T). \end{cases}$$

Close to the critical case ($c \approx 1$), this algorithm becomes extremely slow. Several acceleration methods of the convergence of (P_s) have been introduced and studied. In particular the Diffusion Synthetic Acceleration (DSA) method [12][8] and multigrid algorithms [14][18].

The main difficulties encountered while studying these methods lead the authors either to consider the discretized equation in the angular variable [11][15], or the continuous equation with a truncated expansion of k with respect to this angular variable [18][11].

To our knowledge, the only theoretical proof for the acceleration of the convergence in the continuous case (in space an angular variables) has been obtained for reflexive boundary conditions by [12].

The idea of [2] and [3] is to introduce and study better algorithms than (P_s), adapted from the methods of Jacobi, Gauss-Seidel and SOR, in the infinite dimensional case. These algorithms can be accelerated by an adapted DSA method. This approach has been studied in dimension one and two by [6], and successfully compared to standard DSA method.

Our aim is to propose a new algorithm, replacing Jacobi, Gauss-Seidel or SOR algorithms, based on an adaptation of the minimal residual method in infinite dimensional case. As others algorithms, it relies on the following natural splitting of k.

Let K_{ij}, $i,j \in \{1,\dots,4\}$ be the integral operator whose kernel is

$$k_{ij}(x,\Omega,\Omega') = k(x,\Omega,\Omega') \times \mathbf{1}_{Q_i}(x,\Omega) \times \mathbf{1}_{Q_j}(x,\Omega'),$$

with $Q_i = D \times B_i$, B_i being the i-th quarter of the unit disk (see figure 1) and $\mathbf{1}_{Q_i}(x,\Omega)$ the indicator function of Q_i.

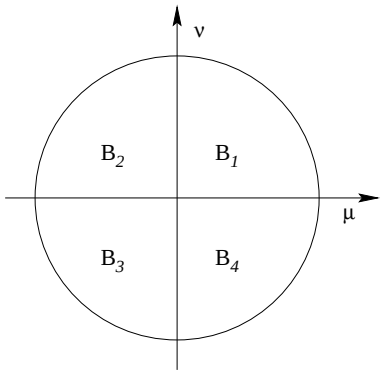

FIG. 1. *Decomposition of B*

Since we have $K_{ij}(f) = K_{ij}(f.\mathbf{1}_{Q_j})\mathbf{1}_{Q_i}$, operator K splits into $K = \sum_{i,j=1}^{4} K_{ij}$. Note that K_{ij} is an operator acting from $L^2(Q)$, using only the values of f on Q_j, such that $K_{ij}f$ has its support in Q_i. The solution of (P) is given by $f = f_1 + f_2 + f_3 + f_4$ with $f_1, f_2, f_3, f_4 \in \mathcal{D}(T)$ solution of

$$\begin{pmatrix} T-K_{11} & -K_{12} & -K_{13} & -K_{14} \\ -K_{21} & T-K_{22} & -K_{23} & -K_{24} \\ -K_{31} & -K_{32} & T-K_{33} & -K_{34} \\ -K_{41} & -K_{42} & -K_{43} & T-K_{44} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} = \begin{pmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{pmatrix} \tag{1}$$

where $S_i = S \times \mathbf{1}_{Q_i}$. Then we have $f_i = f \times \mathbf{1}_{Q_i}$ for $i \in \{1,\dots,4\}$. The SOR method introduced by [6] gives excellent results, but needs the computation of its optimal parameter, which in turn can be very slow in the critical case. For these reasons we looked for a method that gives good rate of convergence, but do not need any extra parameter calculation.

2. Minimal residual algorithm. This method was introduced by O. Axelsson [13], in the finite dimensional case, and proved to converge provided the matrix of the linear system has a definite positive symmetric part.
Using the operator splitting devised by S. Akesbi and M. Nicolet, the transport equa-

tion is equivalent to the following system

$$\begin{pmatrix} I-\theta_{11} & -\theta_{12} & -\theta_{13} & -\theta_{14} \\ -\theta_{21} & I-\theta_{22} & -\theta_{23} & -\theta_{24} \\ -\theta_{31} & -\theta_{32} & I-\theta_{33} & -\theta_{34} \\ -\theta_{41} & -\theta_{42} & -\theta_{43} & I-\theta_{44} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} = \begin{pmatrix} \widetilde{S_1} \\ \widetilde{S_2} \\ \widetilde{S_3} \\ \widetilde{S_4} \end{pmatrix},$$

where we applied on components the operator T^{-1}, and set $\theta_{ij} = T^{-1}K_{ij}$, $\widetilde{S}_i = T^{-1}S_i$. The matrix of operators of our system will be preconditioned by the inverse of diagonal i.e.

$$\begin{pmatrix} (I-\theta_{11})^{-1} & 0 & 0 & 0 \\ 0 & (I-\theta_{22})^{-1} & 0 & 0 \\ 0 & 0 & (I-\theta_{33})^{-1} & 0 \\ 0 & 0 & 0 & (I-\theta_{44})^{-1} \end{pmatrix},$$

leading to the following matrix of operators

$$\mathcal{A} = \begin{pmatrix} I & -(I-\theta_{11})^{-1}\theta_{12} & -(I-\theta_{11})^{-1}\theta_{13} & -(I-\theta_{11})^{-1}\theta_{14} \\ -(I-\theta_{22})^{-1}\theta_{21} & I & -(I-\theta_{22})^{-1}\theta_{23} & -(I-\theta_{22})^{-1}\theta_{24} \\ -(I-\theta_{33})^{-1}\theta_{31} & -(I-\theta_{33})^{-1}\theta_{32} & I & -(I-\theta_{33})^{-1}\theta_{34} \\ -(I-\theta_{44})^{-1}\theta_{41} & -(I-\theta_{44})^{-1}\theta_{42} & -(I-\theta_{44})^{-1}\theta_{43} & I \end{pmatrix}$$

In order to perform a minimal residual method, we have to make clear which operations between matrix and vectors, appearing in the method, can be calculated from a numerical point of view.

We are willing to solve $\mathcal{A}F = B$, where $F = {}^t(f_1, f_2, f_3, f_4) \in \mathcal{D}(T)^4$. We denote by $\langle , \rangle$ the scalar product in $(L^2(Q))^4$, i.e. $\langle F, G\rangle = (f_1, g_1) + (f_2, g_2) + (f_3, g_3) + (f_4, g_4)$ where $(,)$ is the standard $L^2(Q)$ scalar product. Similarly, $\|\|_2$ will represent the norm in $(L^2(Q))^4$ associated to this scalar product.

The minimal residual method, minimizing $\mathcal{E}(F) = \|B - \mathcal{A}F\|_2^2$, takes the following form :

Let $f^0 \in \mathcal{D}(T)$, $F^0 = (f^0 \mathbf{1}_{Q_i})_{i=1,\dots,4}$, $R^0 = B - \mathcal{A}F^0$, $P^0 = R^0$, $Q^0 = \mathcal{A}P^0$.

While $\left\|R^k\right\|_2 > \varepsilon$ do

begin

$$\begin{aligned} \alpha^k &= \frac{\langle R^k, Q^k\rangle}{\langle Q^k, Q^k\rangle} \\ F^{k+1} &= F^k + \alpha^k P^k \\ R^{k+1} &= R^k - \alpha^k Q^k \\ \beta^{k+1} &= -\frac{\langle \mathcal{A}R^{k+1}, Q^k\rangle}{\langle Q^k, Q^k\rangle} \\ P^{k+1} &= R^{k+1} + \beta^{k+1} P^k \\ Q^{k+1} &= \mathcal{A}R^{k+1} + \beta^{k+1} Q^k \end{aligned}$$

end

In the previous algorithm, we have to make clear how we compute the product $\mathcal{A}$ times a vector, since $\mathcal{A}$ contains some inverse operator.

So let $g \in \mathcal{D}(T)$, $\mathcal{G} = (g\mathbf{1}_{Q_i})_{i=1,\dots,4}$ and see how to compute $\mathcal{Z} = (z_1, z_2, z_3, z_4)$ verifying

$$\mathcal{Z} = \mathcal{A}\mathcal{G}$$

Componentwise, this equality means for $i = 1, \dots, 4$,

$$z_i = g_i - \sum_{j \neq i} (I - \theta_{ii})^{-1} \theta_{ij} g_j.$$

Applying $T(I - \theta_{ii}) = T - K_{ii}$ to the first equation we get

$$(T - K_{ii})(g_i - z_i) = \sum_{j \neq i} K_{ij} g_j \tag{2}$$

These integro-differential equations can be calculated numerically [1] thanks to the splitting and the special form of the kernel assumed in (A4). More explicitly, the equation $R^0 = B - \mathcal{A}F^0$ corresponds to solve the system

$$(T - K_{ii})(r_i^0 + f_i^0) = S_i + \sum_{j \neq i} K_{ij} f_j^0, \qquad i = 1, \dots, 4,$$

whereas $Q^0 = \mathcal{A}P^0$ stands for

$$(T - K_{ii})(p_i^0 - q_i^0) = \sum_{j \neq i} K_{ij} p_j^0, \qquad i = 1, \dots, 4.$$

At last, the product $\mathcal{A}R^{k+1} =: D^{k+1}$ which of course is calculated only one time per iteration, is associated to the following equations :

$$(T - K_{ii})(r_i^{k+1} - d_i^{k+1}) = \sum_{j \neq i} K_{ij} r_j^{k+1}, \qquad i = 1, \dots, 4.$$

REMARK 2. *Equations (2) correspond to one step of a Jacobi iteration. One could also thing of a Gauss-Seidel iteration, which would be in that case*

$$(T - K_{ii})(g_i - z_i) = \sum_{j > i} K_{ij} g_j - \sum_{j < i} K_{ij} z_j.$$

Of course one may also perform a symmetric Gauss-Seidel iteration; in what follows we study the convergence of this iterative method with a Jacobi type iteration. We present numerical results for Jacobi, Gauss-Seidel and symmetric Gauss-Seidel iterations.

2.1. Rate of residual decreasing. Applying elementary analysis of [13], we have the following estimate on the residual (cf [4] for the proof) :

PROPOSITION 1. *Let F^k be constructed by the preceding algorithm starting from F^0. Then for $k \geq 0$,*

$$\mathcal{E}(F^{k+1}) \leq \mathcal{E}(F^k) \left(1 - \frac{\langle R^k, \mathcal{A}R^k \rangle}{\langle R^k, R^k \rangle} \frac{\langle R^k, \mathcal{A}R^k \rangle}{\langle \mathcal{A}R^k, \mathcal{A}R^k \rangle} \right). \tag{3}$$

2.2. Theoretical convergence. Let us prove first that our operator $\mathcal{A}$ has somehow definite positive symmetric part, so that for some $\lambda > 0$,

$$\frac{\langle R^k, \mathcal{A}R^k \rangle}{\langle R^k, R^k \rangle} \geq \lambda.$$

PROPOSITION 2. *Under assumption (A1)-(A3) the operator $\mathcal{A}$ has a definite positive symmetric part and verifies*

$$\langle \mathcal{A}F, F \rangle \geq \frac{1-c}{1-\frac{c}{4}} \|F\|_2^2, \qquad \forall F \in \mathcal{D}(T)^4. \tag{4}$$

Proof. We have

$$\begin{aligned} \langle \mathcal{A}F, F \rangle &= \sum_{i=1}^{4} \|F_i\|_2^2 - \sum_{i=1}^{4} \sum_{j \neq i} (F_i, (I-\theta_{ii})^{-1} \theta_{ij} F_j) \\ &\geq \sum_{i=1}^{4} \|F_i\|_2^2 - \gamma \sum_{i=1}^{4} \sum_{j \neq i} \|F_i\|_2 \|F_j\|_2 \end{aligned}$$

if $\|(I-\theta_{ii})^{-1}\theta_{ij}\|_2 \leq \gamma$. The corresponding symmetric bilinear form on $\mathbf{R}^4$ is definite positive when $\gamma < \frac{1}{3}$, since its eigenvalues are $1-3\gamma$ and $1+\gamma$. It remains to control the norms of θ_{ii} and θ_{ij}. To this end, let us state the following

LEMMA 3. *Under assumptions (A1)-(A3), $\|\theta_{ij}\|_2 \leq \frac{c}{4}$, for $(i,j) \in \{1,\ldots,4\}^2$.*

Postponing the demonstration of this lemma, we compute

$$\begin{aligned} \|(I-\theta_{ii})^{-1}\theta_{ij}\|_2 &= \sup_{\|f\|_2=1} \|(I-\theta_{ii})^{-1}\theta_{ij} f\|_2 \\ &= \sup_{\|f\|_2=1} \left\| \left(\sum_{k=0}^{\infty} \theta_{ii}^k \right) \theta_{ij} f \right\|_2 \\ &\leq \left(\sum_{k=0}^{\infty} \|\theta_{ii}\|_2^k \right) \|\theta_{ij}\|_2 \\ &\leq \frac{\|\theta_{ij}\|_2}{1-\|\theta_{ii}\|_2}. \end{aligned}$$

Thus from lemma

$$\|(I-\theta_{ii})^{-1}\theta_{ij}\|_2 \leq \frac{\frac{c}{4}}{1-\frac{c}{4}} = \frac{c}{4-c}.$$

Therefore taking $\gamma = \frac{c}{4-c} < \frac{1}{3}$ as $c < 1$, the smallest eigenvalue of the bilinear form on $\mathbf{R}^4$ is

$$1 - 3\gamma = \frac{1-c}{1-\frac{c}{4}}.$$

The result follows. □

Proof. [of lemma 1.] Recall that $A = \Omega \cdot \nabla_x$ is m-accretive on $\mathcal{D}(T)$. It induces a m-accretive operator A_i on $L^2(Q_i)$ whose domain is

$$\mathcal{D}(A_i) = \left\{ f \in L^2(Q_i) \ : \ \Omega \cdot \nabla_x f \in L^2(Q_i), \ f = 0 \text{ on } \Gamma_i^- \right\}$$

where $\Gamma_i^- = \Gamma^- \cap (\partial D \times B_i)$. Let $g \in \mathcal{D}(A_i)$ solution of $Tg = K_{ij}f$ with $f \in L^2(Q_j)$. Then

$$(Tg, g) = (A_i g, g) + (\sigma g, g) \geq (\sigma g, g).$$

Thus

$$(\sigma g, g) \leq (K_{ij} f, g).$$

But thanks to Cauchy-Schwarz inequality,

$$\begin{aligned}
(K_{ij}f, g) &= \int_D \int_{B_i \times B_j} k(x, \Omega, \Omega') f(x, \Omega') g(x, \Omega) d\Omega' d\Omega dx \\
&\leq \int_D \left(\int_{B_i \times B_j} k(x, \Omega, \Omega') f(x, \Omega')^2 d\Omega' d\Omega \right)^{\frac{1}{2}} \left(\int_{B_i \times B_j} k(x, \Omega, \Omega') g(x, \Omega)^2 d\Omega' d\Omega \right)^{\frac{1}{2}} dx \\
&= \int_D \left(\int_{B_j} f(x, \Omega')^2 \left[\int_{B_i} k(x, \Omega, \Omega') d\Omega \right] d\Omega' \right)^{\frac{1}{2}} \\
&\qquad \times \left(\int_{B_i} g(x, \Omega)^2 \left[\int_{B_j} k(x, \Omega, \Omega') d\Omega' \right] d\Omega \right)^{\frac{1}{2}} dx.
\end{aligned}$$

Observe that assumption (A2) on k imply that condition (A3) reads

$$\int_{B_i} k(x, \Omega, \Omega') d\Omega \leq \frac{\sigma_0 c}{4} \qquad \text{and} \qquad \int_{B_j} k(x, \Omega, \Omega') d\Omega' \leq \frac{\sigma_0 c}{4}$$

Finally,

$$\sigma_0 \|g\|_2^2 \leq (\sigma g, g) \leq (K_{ij} f, g) \leq \frac{\sigma_0 c}{4} \|f\|_2 \|g\|_2,$$

the last two norms being taken on Q_j and Q_i respectively. This gives the announced bound. □

Now we turn to the second expression appearing in (3), to prove that for some $\nu > 0$ we have

$$\frac{\langle R^k, \mathcal{A} R^k \rangle}{\langle \mathcal{A} R^k, \mathcal{A} R^k \rangle} \geq \nu.$$

In fact we can give an explicit value for ν :

PROPOSITION 4. *Under assumptions (A1)-(A3), the matrix (of operators) $\mathcal{A}$ verifies*

$$\langle \mathcal{A}F, F \rangle \geq \frac{4 - c}{2(2 + c)} \langle \mathcal{A}F, \mathcal{A}F \rangle, \qquad \forall F \in \mathcal{D}(T)^4. \tag{5}$$

Proof. Denoting by $J_{ij} = (I - \theta_{ii})^{-1} \theta_{ij}$, we have

$$\begin{aligned}
\langle \mathcal{A}F, \mathcal{A}F \rangle &= \sum_{i=1}^4 \left\| F_i - \sum_{j \neq i} J_{ij} F_j \right\|_2^2 \\
&= \sum_{i=1}^4 \left[\|F_i\|_2^2 + \sum_{j \neq i, k \neq i} (J_{ij} F_j, J_{ik} F_k) - 2 \sum_{j \neq i} (J_{ij} F_j, F_i) \right]
\end{aligned}$$

and

$$\langle \mathcal{A}F, F\rangle = \sum_{i=1}^{4}\left[\|F_i\|_2^2 - \sum_{j\neq i}(J_{ij}F_j, F_i)\right].$$

so that for $\nu > \frac{1}{2}$ (we hope we would find such a ν),

$$\begin{aligned}\langle \mathcal{A}F, F\rangle - \nu\,\langle \mathcal{A}F, \mathcal{A}F\rangle &= \sum_{i=1}^{4}\left[(1-\nu)\,\|F_i\|_2^2 - \nu \sum_{j\neq i,k\neq i}(J_{ij}F_j, J_{ik}F_k)\right.\\ &\qquad\left. +\ (2\nu-1)\sum_{j\neq i}(J_{ij}F_j, F_i)\right]\\ &\geq \sum_{i=1}^{4}\left[(1-\nu)\,\|F_i\|_2^2 - \nu\left(\frac{c}{4-c}\right)^2 \sum_{j\neq i,k\neq i}\|F_j\|_2\,\|F_k\|_2\right.\\ &\qquad\left. -(2\nu-1)\frac{c}{4-c}\sum_{j\neq i}\|F_i\|_2\,\|F_j\|_2\right]\end{aligned}$$

since we showed in the proof of previous proposition that $\|J_{ij}\|_2 \leq \frac{c}{4-c}$. Once again, we consider the associated symmetric bilinear form on $\mathbf{R}^4$, which after gathering terms is

$$\begin{aligned}\Phi(X) = \sum_{i=1}^{4}\left[\left(1-\nu-3\nu\left(\frac{c}{4-c}\right)^2\right)X_i^2\right.\\ \left. -\left((2\nu-1)\frac{c}{4-c}+2\nu\left(\frac{c}{4-c}\right)^2\right)\sum_{j\neq i}X_iX_j\right]\end{aligned}$$

It is clear that for this bilinear form is positive if and only if

$$1-\nu-3\nu\left(\frac{c}{4-c}\right)^2 \geq 3(2\nu-1)\frac{c}{4-c}+6\nu\left(\frac{c}{4-c}\right)^2$$

which gives

$$\nu \leq \frac{(4-c)}{2(2+c)}.$$

We easily verify that this value is always greater that one half (see figure 2). □

We can now state the convergence result.

THEOREM 5. *Under assumptions (A1)-(A3), the minimal residual method converges, i.e. F^k converges toward the unique solution of (1), and the residual decreases at least at the following rate :*

$$\mathcal{E}(F^{k+1}) \leq \mathcal{E}(F^k)\left(1-\frac{1-c}{1+\frac{c}{2}}\right) \qquad \text{for } k \geq 0. \tag{6}$$

Proof. Plugging estimations (4) and (5) into (3), we get (6). As $c < 1$, this means that $\mathcal{E}(F^{k+1})$ converges toward 0 when k goes to infinity. Using (4) we have

$$\|F^{k+1} - \mathcal{A}^{-1}B\|_2^2 \leq \frac{(4-c)}{4(1-c)} \left\langle \mathcal{A}F^{k+1} - B, F^{k+1} - \mathcal{A}^{-1}B \right\rangle$$

so that $\|F^{k+1} - \mathcal{A}^{-1}B\|_2 \leq \frac{(4-c)}{4(1-c)} \mathcal{E}(F^{k+1})^{\frac{1}{2}}$ which means $F^{k+1} \to F$ where $\mathcal{A}F = B$. □

REMARK 3. *Our estimate of the convergence rate (6) is not optimal. Indeed, the forthcoming numerical tests will show that our algorithm works for values of c greater than one (see figures 7 and 8).*

REMARK 4. *We see from (6) that convergence is ensured if $c < 1$. In the trivial case when $c = 0$, we find that our method converges in one iteration, since $\mathcal{E}(F^1) = 0$. We could expect this since in this case, there is no coupling between components ($\mathcal{A}$ is the identity). We draw on figure 2 the behavior of the three constants appearing in (4)(5) and (6).*

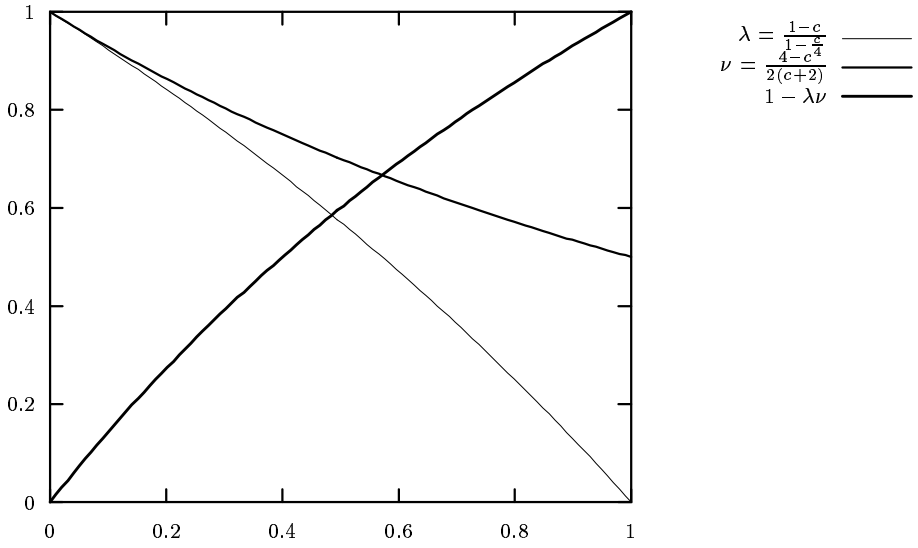

FIG. 2. *Constants appearing in (4)(5) and (6)*

3. Discretization. Let $D = (0, a) \times (0, b)$. We consider the following triangulation of D :

$$\overline{D} = \bigcup_{i,j} ([x_i, x_{i+1}] \times [y_j, y_{j+1}]) = \bigcup_{i,j} D_{i,j}$$

and a triangulation of the disk B : $\overline{B} = \bigcup_k T_k$. Every iteration of our algorithm relies on the resolution of the following problem :

$$\begin{cases} (T - K_{ll}) f_l = g, \\ f_l \in D(T) \end{cases}$$

where g has is support included in Q_l, and $l \in \{1,2,3,4\}$. Without loss of generality we can consider the case $l = 1$. For a kernel k respecting assumption (A4), our problem becomes :

$$\mu\frac{\partial f_1}{\partial x} + \eta\frac{\partial f_1}{\partial y} + \sigma f_1 = C(x,y)\sum_{l=1}^{N_k}\alpha_l(\mu,\eta)\int_{B_1}\alpha_l(\mu',\eta')f_1(x,y,\mu',\eta')d\mu'd\eta' + g$$

We call $P_h(k,l)$ the set of functions $f_h(x,y,\mu,\eta)$ defined on Q such that their restriction to $D_{i,j}\times T_k$ is a polynomial of degree less or equal to k with respect to the spatial coordinates x,y and of degree less or equal to l with respect to the angular variables μ,η. We introduce the discrete space V_h as the space of functions $f_h \in P_h(1,0)$ vanishing on Γ^-, such that

$$x \to \int_{y_j}^{y_{j+1}}\int_{T_k} f_h(x,y,\mu,\eta)dyd\mu d\eta \text{ and } y \to \int_{x_i}^{x_{i+1}}\int_{T_k} f_h(x,y,\mu,\eta)dxd\mu d\eta$$

are continuous functions. We set

$$m_{i,j,k} = \frac{1}{|D_{i,j}|\times|T_k|}\int_{D_{i,j}\times T_k} f_h(x,y,\mu,\eta)dxdyd\mu d\eta$$

$$\Gamma^x_{i,j,k} = \frac{1}{h_x\times|T_k|}\int_{x_i}^{x_{i+1}}\int_{T_k} f_h(x,y_j,\mu,\eta)dxd\mu d\eta$$

$$\Gamma^y_{i,j,k} = \frac{1}{h_y\times|T_k|}\int_{y_j}^{y_{j+1}}\int_{T_k} f_h(x_i,y,\mu,\eta)dyd\mu d\eta$$

where $h_x = x_{i+1} - x_i$ and $h_y = y_{j+1} - y_j$. We denote by π_h the projector from $D(T)$ on $P_h(0,0)$:

$$\pi_h(f)_{|D_{i,j}\times T_k} = \frac{1}{|D_{i,j}|\times|T_k|}\int_{D_{i,j}\times T_k} f(x,y,\mu,\eta)dxdyd\mu d\eta.$$

Taking into account assumption (A4), we define the discrete operator A_h on $D(T)$ by

$$A_h(f) = \pi_h(\mu\frac{\partial f}{\partial x} + \eta\frac{\partial f}{\partial y}) + \pi_h(\sigma)\pi_h(f) - \pi_h(C)\sum_{l=1}^{N_k}\pi_h(\alpha_l)(\int_{B_1}\pi_h(\alpha_l)(\mu',\eta')\pi_h(f)(x,y,\mu',\eta')d\mu'd\eta')$$

and we consider the associated problem :

$$(P_h)\begin{cases}\text{Find } f_h \in V_h \text{ such that}\\ A_h(f_h) = \pi_h(g)\end{cases}$$

Observe that for each $f_h \in V_h$ the imposed continuity conditions lead to :

$$m_{i,j,k} = \frac{1}{2}(\Gamma^x_{i,j+1,k} + \Gamma^x_{i,j,k}) = \frac{1}{2}(\Gamma^y_{i+1,j,k} + \Gamma^y_{i,j,k}). \tag{7}$$

The discrete problem (P_h) can be written as follows :

$$\left(\frac{2\mu_k}{h_x}+\frac{2\eta_k}{h_y}+\sigma_{i,j}\right)m_{i,j,k}=C_{i,j}\sum_{l=1}^{N_k}\alpha_{l,k}\Phi^l_{i,j}+\frac{2\mu_k}{h_x}\Gamma^y_{i,j,k}+\frac{2\eta_k}{h_y}\Gamma^x_{i,j,k}+g_{i,j,k}, \quad (8)$$

for all i,j and k such that $T_k \in B_1$, where

$$\mu_k=\frac{1}{|T_k|}\int_{T_k}\mu d\mu d\eta, \quad \eta_k=\frac{1}{|T_k|}\int_{T_k}\eta d\mu d\eta, \quad \sigma_{i,j}=\frac{1}{|D_{i,j}|}\int_{D_{i,j}}\sigma(x,y)dxdy,$$

$$C_{i,j}=\frac{1}{|D_{i,j}|}\int_{D_{i,j}}C(x,y)dxdy, \quad g_{i,j,k}=\frac{1}{|D_{i,j}|\times|T_k|}\int_{D_{i,j}\times T_k}g(x,y,\mu,\eta)dxdyd\mu d\eta$$

are known quantities, and $\Phi^l_{i,j}=\sum_{k'/T_{k'}\in B_1}\alpha_{l,k'}m_{i,j,k'}$ is unknown.

The incoming fluxes are given on Q_1 by $\Gamma^y_{0,j,k}=\Gamma^x_{i,0,k}=0$. We now explain how to compute $\Gamma^y_{i+1,j,k}$ and $\Gamma^x_{i,j+1,k}$ from $\Gamma^y_{i,j,k}$ and $\Gamma^x_{i,j,k}$.

Multiplying (8) by $\left(\frac{2\mu_k}{h_x}+\frac{2\eta_k}{h_y}+\sigma_{i,j}\right)^{-1}\alpha_{l',k}$ and summing on k, we obtain for each $l' \in \{1,2,...,N_k\}$ a linear equation between the unknown quantities $\Phi^l_{i,j}$, which leads to a small linear system $N_k \times N_k$. Once this system has been solved, the $\Phi^l_{i,j}$ are used in (8) to compute $m_{i,j,k}$ which in turn are plugged into (7) to get $\Gamma^x_{i,j+1,k}$ and $\Gamma^y_{i+1,j,k}$.

For the presented numerical results, we consider $D=(0,1)\times(0,1)$ and $h_x=h_y=\frac{1}{10}$. Each quarter of the unit disk is subdivided into 25 mesh elements. We take a constant kernel $k(\Omega,\Omega')=\frac{\sigma c}{\pi}$. The exact solution of our test problem is given by

$$f(x,y,\mu,\eta)=\begin{cases} xy \text{ on } Q_1 \\ (1-x)y \text{ on } Q_2 \\ (1-x)(1-y) \text{ on } Q_3 \\ x(1-y) \text{ on } Q_4 \end{cases}$$

For every iterative methods tested there, iterations are stopped when $\frac{\left\|F^{k+1}-F^k\right\|_1}{\|F^{k+1}\|_1}$ is less than a prescribed $\varepsilon>0$.

4. Numerical results and discussion. We compare our methods with Gauss-Seidel method and SOR, which has been proved to be a very efficient method. We have to keep in mind that SOR needs the computation of a relaxation parameter which is very time consuming : in one dimension a formula exists for this relaxation parameter, which needs the computation of the spectral radius of Jacobi iterations. In dimension two there is no known formula for this parameter, thus it should be determined by dichotomy. We did not include computation time for this parameter in SOR in all forthcoming tests.

There is two sets of tests : one at fixed σ, another for fixed c. For each case, we first compare all the methods, and then we remove Gauss-Seidel to compare the methods for critical values : c near one and large σ. As shown in figure 3, our method with symmetric Gauss-Seidel solver seems very close to SOR, without computation of any optimal parameter. For values of c close to 1, the situation is even better since

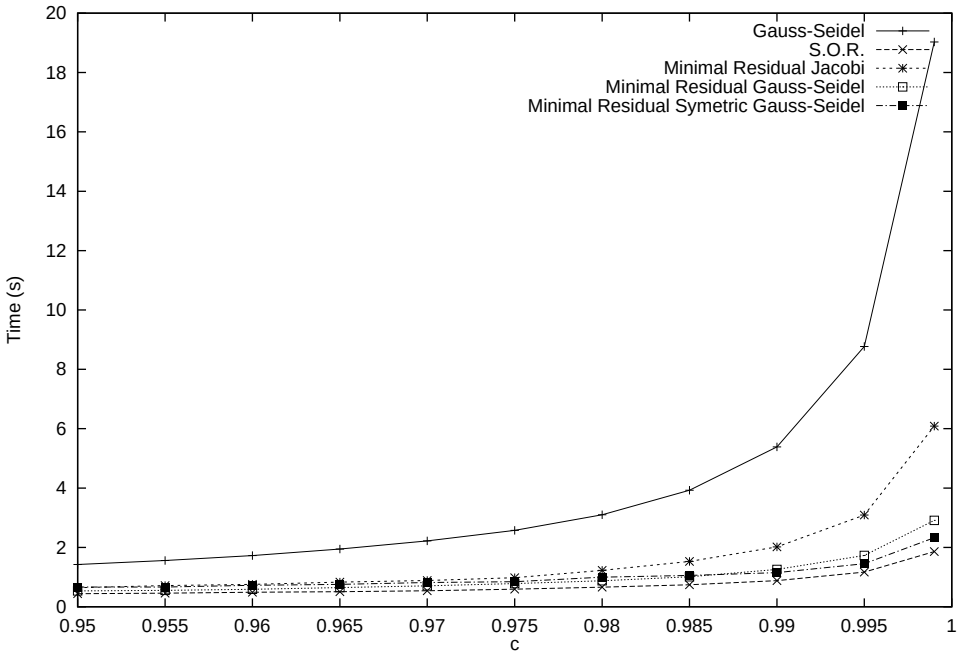

FIG. 3. *Comparison of cpu time at fixed* $\sigma = 50$.

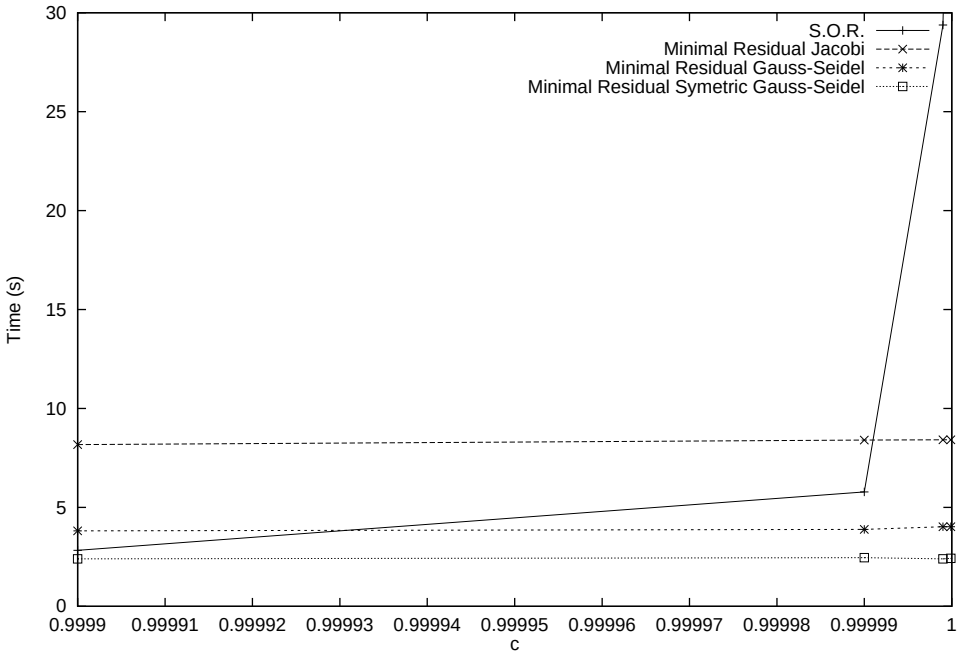

FIG. 4. *Comparison of cpu time at fixed* $\sigma = 50$ *near* $c = 1$.

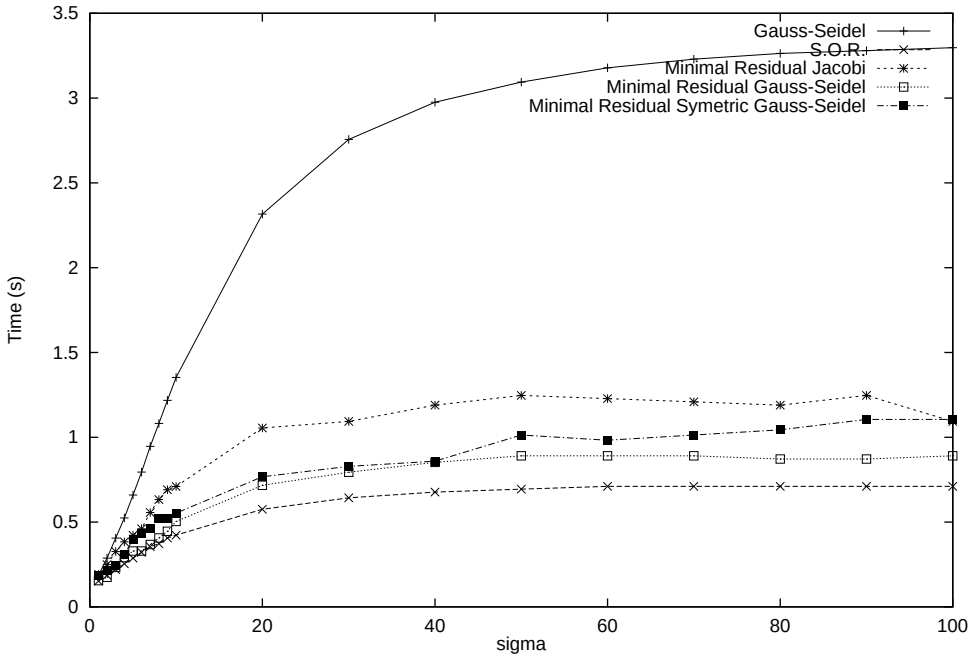

FIG. 5. *Comparison of cpu time at fixed* $c = 0.98$.

SOR converges in more and more iterations near $c = 1$, whereas we can compute the solution for $c = 1$ with our method, as seen on figure 4.

Turning now to the σ dependence, you could see on figure 5 that our schemes are comparable to SOR (again in which we did not count the time spent to compute the relaxation parameter) for small values of σ. A test for really large values of σ reveals that our scheme converges more and more rapidly as σ increase, whereas SOR keeps a constant number of iterations (see figure 6 in decimal log scale).

As our algorithm seemed to converge even for $c > 1$, we plotted in figures 7 and 8 some tests for c from 1 to 4. This last value is a critical value for which our numerical method may fail to work (the leading coefficient in (8) may vanish). Note that our algorithm is still more efficient for great values of σ.

5. Conclusions. We showed through the previous numerical tests that our methods are as efficient as SOR for non-critical cases (c close to 1 or large σ), and converge even faster for critical cases. Moreover, their implementation is as easy as standard algorithm (P_s). They are naturally devised for parallelization. A work is in progress for the acceleration of this algorithm by an adapted DSA method [6], and its comparison with standard DSA.

We noticed during our numerical tests a faster rate of convergence of our algorithms than theoretically estimated by (6). Maybe this estimation could be improved. []

REFERENCES

[1] S. AKESBI, *Splitting d'opérateur pour l'équation de transport neutronique en géométrie bidimensionnelle plane.* submitted.

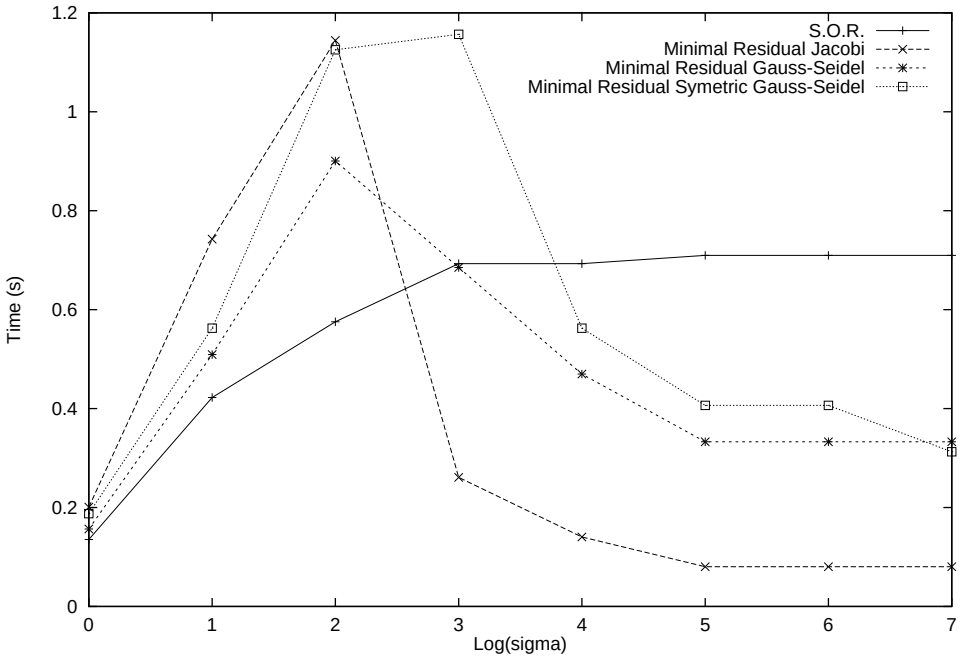

FIG. 6. *Comparison of cpu time at fixed* $c = 0.98$, *for large* σ.

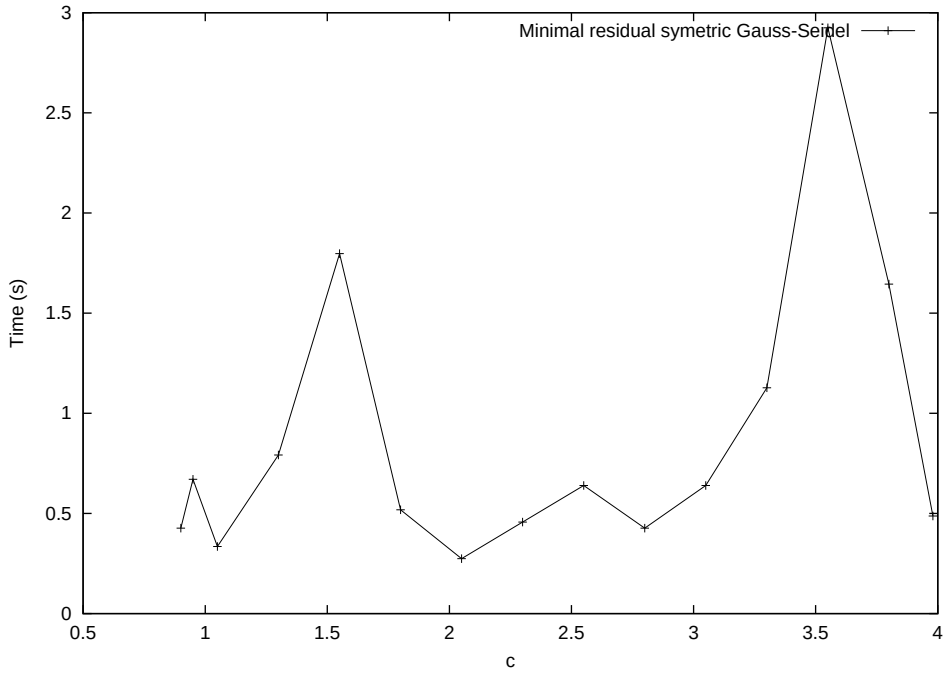

FIG. 7. *Cpu time at fixed* $\sigma = 50$, *for* $c > 1$.

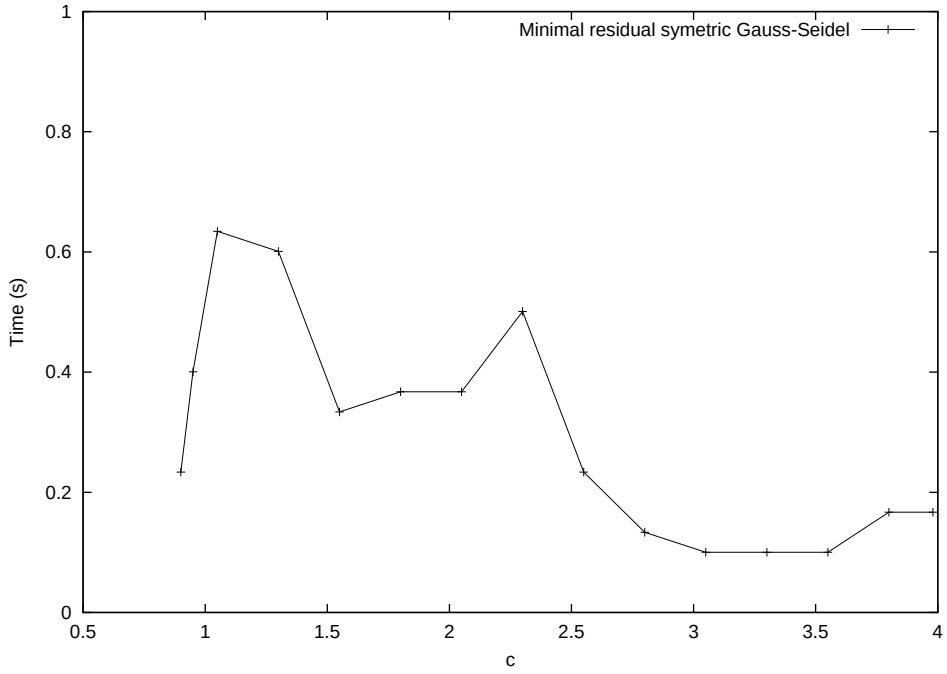

FIG. 8. *Cpu time at fixed* $\sigma = 10^6$*, for* $c > 1$.

[2] S. AKESBI, M. LAYDI, AND M. MOKHTAR-KHARROUBI, *Décomposition d'opérateurs et accélération de la convergence en neutronique*, C.R. Acad. Sci. Paris, 319 (1994), pp. 765–770.
[3] ———, *Schemes and acceleration in transport theory*, To appear in Journ. of Transport Theory and Statistical Physics, (1999).
[4] S. AKESBI AND E. MAITRE, *Minimal residual method applied to transport equation.* submitted.
[5] S. AKESBI AND M. NICOLET, *Nouveaux algorithmes pour l'équation de transport neutronique en gómétrie bidimensionnelle*, C.R. Acad. Sci. Paris t. 324, Série I, (1997), pp. 699–706.
[6] ———, *Nouveaux algorithmes performants en théorie du transport*, M2AN, 32 (1998), pp. 341–358.
[7] ALCOUFFE-CLARK-LARSEN, *The diffusion synthetic acceleration in multiple time scales*, J. Brackbill, editor Ac. Press, 1985.
[8] S. F. ASHBY, P. N. BROWN, M. R. DORR, AND A. C. HINDMARSH, *A linear algebraic analysis of diffusion synthetic acceleration for the boltzmann tranport equation*, SIAM Journal on Numerical Analysis, 32 (1995), pp. 128–178.
[9] J. M. BANOCZI AND C. T. KELLEY, *A fast multilevel algorithm for the solution of nonlinear systems of conductive-radiative heat transfer equations in two spaces dimensions*, SIAM J. Sci. Comp., 20 (1999), pp. 1214–1228.
[10] R. DAUTRAY AND J.-L.LIONS, *Analyse mathématique et calcul numérique*, vol. 9, Masson, 1987.
[11] K. M. KHATTAB AND E. W. LARSEN, *Synthetic acceleration methods for linear transport problems with highly anisotropic scattering*, Nuclear Science and Engineering, 107 (1991), pp. 217–227.
[12] E. LARSEN, *Unconditionally stable diffusion-synthetic acceleration methods for the slab geometry dicrete-ordinates equations*, Nucl. Sci. and Eng., parts I-II (1988).
[13] P. LASCAUX AND R. THÉODOR, *Analyse numérique matricelle appliquée à l'art de l'ingénieur*, vol. 2, Masson, 1987.
[14] T. MANTEUFFEL, S. MCCORMICK, J. MOREL, AND G. YANG, *A fast multigrid algorithm for isotropic transport problems ii. with absorption*, SIAM Journal on Scientific Computing, 17 (1996), pp. 1449–1474.
[15] T. MANTEUFFEL AND K. RESSEL, *Least-squares finite-element solution for the neutronic transport equation in diffusive regimes*, SIAM Journal on Numerical Analysis, 35 (1998), pp. 806–835.

[16] I. Marek, *Frobenius theory of positive operators, comparison theorems and applications*, Siam. Journ. Appl. Math., 19 (1970).
[17] M. Mokhtar-Kharroubi, *On the approximation of a class of transport equations*, Transport Theory and Statistical Physics, 22 (1993), pp. 561–570.
[18] J. E. Morel and T. A. Manteuffel, *An angular multigrid acceleration technique for s_n equations with highly forward-peaked scattering*, Nuclear Science and Engineering, 107 (1991), pp. 330–342.
[19] P. Nelson, *A survey convergence results in numerical transport theory*, in Com. Proceedings in honor of G.M. Wing's 65th birthday. Transport Theory, Invariant Embedding, and Integral, E. by P. Nelson and al., eds., 1989.
[20] R. Sanchez and N. J. McCormick, *A review of neutron transport approximations*, NUcl. Sci. and Eng., 80 (1982), pp. 481–535.
[21] R. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Engelwood Cliffs. N.J., 1962.