ESAIM: PROCEEDINGS, July 2007, Vol.18, 70-86
Jean-Frédéric Gerbeau & Stéphane Labbé, Editors

STOCHASTIC MATRICES AND L_p NORMS : NEW ALGORITHMS FOR SOLVING ILL-CONDITIONED LINEAR SYSTEMS OF EQUATIONS

Riadh Zorgati[1, 2], Wim van Ackooij[2] and Marc Lambert[1]

Abstract. We propose new iterative algorithms for solving a system of linear equations, possibly singular and inconsistent, presenting outstanding performances regarding ill-conditioning and error propagation. The basis of our approach is constructing with the l_1 norm, a preconditioning matrix C (an approximation of a generalized inverse of the matrix) such that the preconditioned matrix CA is stochastic. This property allows us to retrieve, in an original way, the Schultz-Hotelling-Bodewig's algorithm of iterative refinement of the approximate inverse of a matrix. The approach, valid for non-negative matrices, is then generalized to any complex, rectangular matrix. We are then able to compute a generalized inverse of any matrix and this inverse is fit for use in classical solving schemes such as : Richardson-Tanabe, Schultz-Hotelling-Bodewig, preconditioned conjugate gradients and also in the Kaczmarz scheme (that we have generalized using l_p norms). Regarding the obtained results on pathological well-known test-cases such as Hilbert and Nakasaka matrices, some of the proposed algorithms are empirically shown to be more efficient than the known classical techniques.

1. Introduction

Consider any system of linear equations:

$$Ax = b, \tag{1.1}$$

where A is a $m \times n$ complex matrix and x and b are complex vectors of dimensions n and m respectively.

When (1.1) is solvable, many methods of resolution can be implemented according to the characteristics of the matrix A. Considering for example, classical methods when A is square, the direct method of gaussian elimination or classical iterative methods as Jacobi or Gauss-Seidel are well-suited for providing precise solutions as long as A has no pathological numerical features. If A is hermitian, positive-definite, the conjugate gradients method of [7] is preferable. For rectangular, possibly singular systems, we can use projective methods of Kaczmarz [11] and Cimmino [5], who converge for any system with nonzero rows to the unique solution of a generalized solution if such solutions exist [1, 18].

[1] Supelec, Plateau de Moulon, 3 rue Joliot-Curie, F-91192 Gif-sur-Yvette Cedex FRANCE. Tel: +33 (0)1 47 65 49 79; e-mail: riadh.zorgati@lss.supelec.fr & marc.lambert@lss.supelec.fr

[2] EDF R&D. 1, avenue du Général de Gaulle, F-92141 Clamart Cedex FRANCE. Tel: +33 (0)1 47 65 58 31; e-mail: riadh.zorgati@edf.fr & wim.van-ackooij@edf.fr

Theoretically, solving (1.1) poses no difficulty. In practice, however it often meets with the snags of error propagation and ill-conditioning of the matrix A. The Gauss method, is particularly sensitive to these two challenges (for illustration of the error propagation snag, see for example [14] and [19]).

As (1.1) may have no classical solution, following [19], we consider solving the system

$$Ax = AA^{-}b, \tag{1.2}$$

instead of (1.1), which always has a solution $x = A^{-}b$, where A^{-} is a generalized inverse of A. Considering (1.2) instead of (1.1) is justified by the fact that when (1.1) is solvable, the sets of solutions of (1.1) and (1.2) coincide due to the equality $AA^{-}b = b$ which holds for any generalized inverse A^{-} of A.

Let us consider the system :

$$RAx = Rb, \tag{1.3}$$

The matrix R is said to be a gain matrix if $Rb \in \mathrm{Im}(RA)$, $\rho(I - RA|_{\mathrm{Im}(I-RA)}) < 1$ where ρ is the spectral radius of the iteration matrix $Q = I - RA|_{\mathrm{Im}(I-RA)}$ and if furthermore $\mathrm{Ker}(RA) = \mathrm{Ker}(A)$. Let $S(H, y)$ be the set of all solutions of equation $Hx = y$. If R is a gain matrix, then equation (1.3) is solvable and the set of solutions of (1.2) coincides with the set of solutions of (1.3), i.e.,

$$S(A, AA^{-}b) = S(RA, Rb).$$

Considering (1.2) instead of (1.1) and then considering (1.3) instead of (1.2) actually results in considering (1.1) preconditioned by a matrix R. The preconditioning matrix R is then chosen in a way to guarantee existence of a solution. Furthermore the choice is such that the computation of the solution from a numerical point of view is tractable.

For a given gain matrix R, three iterative schemes of resolution can be used for solving (1.3). Without considering stopping criteria for simplifying the presentation, these schemes are :

- the Richardson-Tanabe scheme :

```
$x^0 \in \mathrm{C}^n \quad \alpha \in \left]0, \frac{2}{\rho(RA)}\right]$
$For\ k = 1, 2, \ldots$
    $x^{k+1} = x^k + \alpha R\left(b - Ax^k\right)$
$End\ for$
```

  Such linear stationary iterative processes have been characterized by [19].

- the Schultz-Hotelling-Bodewig's scheme of iterative refinement of the inverse of a matrix:

```
$\rho\left(I - RA|_{\mathrm{Im}(I-RA)}\right) < 1 \quad R^0 = R$
$For\ k = 1, 2, \ldots$
    $R^{k+1} = R^k\left(2I - AR^k\right)$
$End\ for$
```

  We will prove that this scheme converges to a generalized inverse G of the matrix A if $\rho(I - RA|_{\mathrm{Im}(I-RA)}) < 1$. A particular solution of (1.3) is then given by $x = Gb$ if $Rb \in \mathrm{Im}(RA)$.

- the Hestenes-Stiefel's conjugate gradients scheme, for any matrix A hermitian, positive-definite :

```
$x_0 \quad p_0 = r_0 = Rb - RAx_0$
$For\ k = 0, 1, ...$
  $\alpha_k = \frac{\|r_k\|^2}{\langle RAp_k, p_k \rangle}$
  $x^{k+1} = x^k + \alpha_k p_k$
  $r^{k+1} = r^k - \alpha_k RAp_k$
  $\beta_{k+1} = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}$
  $p_{k+1} = r^{k+1} + \beta_{k+1} p_k$
$End\ for$
```

Many gain matrices may be chosen. Some of them are quite famous : the Jacobi matrix, defined if all a_{ij} are nonzero and convergent for any diagonally dominant matrix A :

$$R = \begin{pmatrix} \frac{1}{a_{11}} & & 0 \\ & \frac{1}{a_{ij}} & \\ 0 & & \frac{1}{a_{nn}} \end{pmatrix} = \mathrm{diag}(A)^{-1} = J$$

and the matrix $R = \alpha A^*$, built using the adjoint matrix A^*, and convergent for any scalar α satisfying:

$$0 < \alpha < \frac{2}{\rho(A^* A)}$$

This matrix allows us to obtain the Moore-Penrose generalized inverse, called P, and characterized by $APA = A$, $PAP = P$, $(PA)^* = PA$, $(AP)^* = AP$. This means that the Moore-Penrose generalized inverse is a reflexive, a least-squares and a minimum norm generalized inverse.

In addition, we focus on two gain matrices R who are particularly interesting because they always satisfy the convergence condition $\rho(I - RA|_{\mathrm{Im}(I-RA)}) < 1$ for any nonzero row matrix A, as the matrix we will propose in this paper :

- the Kaczmarz-Tanabe matrix, called $K = [K_1 ... K_n]$, with

$$K_i = \frac{1}{\|a_i\|_2^2} \prod_{j=1}^{i-1} (I - \frac{1}{\|a_j\|_2^2} a_j^* a_j) a_i^*,$$

  where the product is considered to be I whenever $i < 2$ and a_i is the ith row-vector of A.
- the Cimmino matrix, called C, expressed as :

$$C = \frac{2}{n} A^* D \quad D = \begin{pmatrix} \frac{1}{\|a_1\|_2^2} & & 0 \\ & . & \\ 0 & & \frac{1}{\|a_n\|_2^2} \end{pmatrix}$$

The Cimmino matrix also allows us to generate, for any initial matrix A, a preconditioned matrix CA hermitian, positive-definite in a way such that the conjugate gradients method can be applied for solving the resulting system.

In practice, we never explicitly compute these two matrices because there are simple iterative schemes for implementing Kaczmarz's and Cimmino's methods (cf. [4] for a complete presentation of the Kaczmarz method).

In this way, any matrix R satisfying the convergence condition $\rho(I - RA|_{\mathrm{Im}(I-RA)}) < 1$ can be considered as an approximation of the inverse of the matrix A. Defining a gain matrix always convergent and having both good theoretical and numerical features is of prime importance for efficiently solving linear systems of equations.

Following this approach, we propose, in this paper, new iterative algorithms for solving any system of linear equations, possibly singular and inconsistent, presenting outstanding performances with respect to ill-conditioning and error propagation. The basis of our approach is constructing with the l_1 norm, a preconditioning matrix C, an approximation of a generalized inverse of the matrix such that the preconditioned matrix CA is a stochastic matrix, i.e. the matrix of states transitions probabilities associated to a stationary Markov chain with n states. This property allows us to retrieve, in a original way, the Schultz-Hotelling-Bodewig's algorithm of iterative refinement of the approximate inverse of a matrix.

This approach, valid for non-negative matrices (i.e. matrices with elements $a_{ij} \geq 0$) is first extended to hermitian, semi-definite-positive matrices and finally generalized to any complex rectangular matrices. We are then able to build a matrix R, $n \times m$, representing a generalized inverse of any matrix A, $m \times n$, always satisfying the convergence condition $\rho(I - RA|_{\mathrm{Im}(I-RA)}) < 1$. Thanks to l_p norms, we show how the Cimmino's matrix can be considered, in our approach, as a particular case : choice of Euclidian norm and asymmetrical structure.

In part 3, we briefly show that the Gauss-Seidel scheme is equivalent to a specific preconditioning of the systems (1.1) or (1.3) using the Jacobi matrix. This original Gauss-Seidel scheme can then be generalized when using any preconditioning matrix, especially with the matrix we suggest.

In part 4, we show how the matrix we propose, thanks to its remarkable properties, can be efficiently used in different solving schemes : Richardson-Tanabe, Schultz-Hotelling-Bodewig, preconditioned conjugate gradients, Kaczmarz generalized by l_p norms. Results on both characterization of the type of generalized inverse obtained and convergence are given.

Hence, we obtain new algorithms with interesting performances, when faced with ill-conditioned matrices and the error propagation. These performances are illustrated in part 5 on some well-known pathologic test-cases including Hilbert matrix for ill-conditioning and Nakasaka tridiagonal matrices for error propagation [14, 18]. We empirically show that some of proposed algorithms are more efficient than the classical techniques we have tested : Gauss, Moore-Penrose inverse, standard and minimum residue, conjugate gradients, Kaczmarz and Cimmino.

We conclude on the main improvements provided by our algorithms and open on a very early prospective application of our approach based on stochastic matrices for computing some parameters of the solution x of (1.1) (as the mean of the components of x, the variance, ...) prior to its resolution. Such an approach, if it were to be efficient, would be an interesting source of information on the solution of a huge, pathological, and, in practice, untractable, system of linear equations.

2. Semi-positive Matrices

In this section, we will present our approach for semi-positive matrices. We will extend this approach in section 3. We will call a matrix semi-positive if all of its elements are larger than or equal to zero and no row or column is identical to zero. We will define a matrix to be positive in a similar way. Throughout this paper, let u denote the all-one vector of appropriate dimension.

Definition 2.1. We denote with $\mathbb{R}_{\geq 0}^{m \times m}$ the space of all semi-positive matrices of dimension m. Let $f : \mathbb{R}_{\geq 0}^{m \times m} \to \mathbb{R}_{\geq 0}^{m \times m}$ be the mapping defined by

$$f(A) = \mathrm{diag}(Au)^{-1}. \tag{2.4}$$

We can easily see that $f(A)A$ is a stochastic matrix (i.e., $f(A)Au = u$), as A is a semi-positive matrix.

Lemma 2.2. *Let S be any matrix such that SA is a stochastic matrix and let x be the solution of the system (1.1). Then x can be decomposed as follows :*

$$x = Sb + (I - SA)\nu,$$

where $\nu = x - \mu u$, a perturbation-vector around x's arithmetic mean μ.

Proof. When we substitute $x = \mu u + \nu$ in equation (1.1) and precondition with S, we obtain :

$$\mu SAu + SA\nu = \mu u + SA\nu = Sb,$$

using the fact that SA is a stochastic matrix. The results follows easily when we add ν to both sides of the inequality. □

Lemma 2.3. *Let A be a semi-positive matrix and suppose that $x = \beta u$ solves the system (1.1) for some $\beta \in \mathbb{R}$. The following scheme :*

$$x^{k+1} = x^k + \alpha f(A)(b - Ax^k), \tag{2.5}$$

initialized with $x^0 = \gamma u$ for some $\gamma \in \mathbb{R}$ will converge in only one iteration, regardless of the properties of A.

Proof. Writing the first iteration step, using the fact that $f(A)A$ is a stochastic matrix and the fact that $x^0 = \gamma u$ easily yields the desired result. □

Using Lemma 2.2, we propose to solve equation (1.1) in the following manner. First we will estimate the mean of x, and at each successive step we will try to improve our estimate of the fluctuation vector ν in order to improve our estimate of x. Doing this gives :

$$\begin{aligned}
x^0 &= Sb \\
\nu^0 &= S(b - \mu Au) \\
x^1 &= Sb + (I - SA)\nu^0 = S(2I - AS)b \\
\nu^1 &= S(2I - AS)(b - \mu Au) \\
x^2 &= S(2I - AS)b + (I - S(2I - AS)A)\nu^1 \\
&= S(2I - AS)[2I - AS(2I - AS)]b,
\end{aligned}$$

It is easily seen that the following recursion emerges

$$\begin{aligned}
h_0(A) &= S \\
h_k(A) &= h_{k-1}(A)(2I - Ah_{k-1}(A)) \\
&= (2I - h_{k-1}(A)A)h_{k-1}(A),
\end{aligned}$$

which is none else than the Schultz-Hotelling-Bodewig scheme [3,9,10,17], with S as initial approximation for the inverse of A (see section 4.2 for more on this scheme).

There is also a link in between our choice for the matrix $S = f(A)$ and Jacobi's matrix. Let us define $r_i = (\sum_{j=1}^{m} \pi_{ij} a_{ij})^{-1}$, and $\tilde{S}(r) = \mathrm{diag}(r)$. Then Jacobi's matrix is obtained, when we choose $\pi_{ij} = \delta_{ij}$, whereas our choice is $\pi_{ij} = 1$.

3. General case

In this paragraph we will extend our method to all matrices, even non-square ones. First of all we will consider hermitian positive definite matrices.

3.1. Positive definite Hermitian matrix

Definition 3.1. We denote with $\mathbb{C}_{+}^{m\times m}$ the space of all complex Hermitian positive definite matrices of dimension m. Let $\hat{f} : \mathbb{C}_{+}^{m\times m} \to \mathbb{R}_{\geq 0}^{m\times m}$ be the mapping defined by

$$\hat{f}(A) = \mathrm{diag}(w(A))^{-1}, \tag{3.6}$$

where w is a mapping that assigns to each matrix A the vector that contains the l_1 norms of each row of A, i.e., $w(A)_i = \|a_i\|_1$.

We note that the above mapping is well defined as A is positive definite and therefore $w(A)_i \neq 0$ for all i.

Lemma 3.2. *If $A \in \mathbb{C}_{+}^{m\times m}$, then the eigenvalues of $\hat{f}(A)A$ are contained in the interval $(0, 1]$.*

Proof. We remark that $\hat{f}(A)A$ can be written as PDP^T since it is the product of positive definite matrix $\hat{f}(A)$ and an Hermitian matrix A (see [8]). Therefore $\hat{f}(A)A$ has the same number of positive, negative and zero eigenvalues as A, which is a positive definite matrix and has only strictly positive eigenvalues. Using the Gershgörin-Hadamard theorem we obtain :

$$\begin{aligned} \lambda_{max}(\hat{f}(A)A) &= \rho(\hat{f}(A)A) \leq \max(\left|\hat{f}(A)A\right| u) \\ &= \max(\hat{f}(A)\,|A|\,u), \end{aligned} \tag{3.7}$$

here the absolute value of a matrix has to be interpreted elements-wise. We remark that $\hat{f}(A)\,|A|$ is a stochastic matrix and therefore the right-hand side of (3.7) is equal to one. □

Lemma 3.3. *If $A \in \mathbb{C}_{+}^{m\times m}$, then the spectral radius of the matrix $Q = I - \alpha\hat{f}(A)A$ is strictly inferior to 1 for any $\alpha \in (0, 2)$.*

Proof. We note that, $\rho(I - \hat{f}(A)A) = \max(1 - \lambda(\hat{f}(A)A)) < 1$, as the above Lemma shows that $\lambda(\hat{f}(A)A)) > 0$. We know furthermore that [12] if $\rho(I - B) < 1$, then $\rho(I - \alpha B) < 1$ for all $0 < \alpha < 2/\rho(B)$, hence the result follows. □

When we combine the above two lemma's we see that the following holds :

Theorem 3.4. *If $A \in \mathbb{C}_{+}^{m\times m}$ and X is a positive definite matrix such that $|XA|$ is a stochastic matrix, then the iteration matrix $Q = I - \alpha XA$ is convergent*[*] *for any $\alpha \in (0, 2)$.*

$\mathrm{Ker}(A\hat{f}(A))$ and $\mathrm{Im}(\hat{f}(A)A)$ determine the nature of the generalized inverse of A when using matrix $\hat{f}(A)$.

Lemma 3.5. *For a regular hermitian, positive definite matrix A, the generalized inverse, called $\tilde{G}$ generated by matrix $R = \hat{f}(A)$, is reflexive, of A^*-minimal-norm and $\hat{f}(A)$-least squares.*

Proof. By application of Theorem 16 [Tanabe, 1974]. □

[*] i.e., its spectral radius is strictly inferior to 1.

3.2. **Any matrix**

We will extend the previous definition to any matrix, even those which are not square.

Definition 3.6. We denote with $\mathbb{C}^{m\times n}_{\neq 0}$ the space of all complex matrices of size $m \times n$, with no zero row or column. Let $f_{mn} : \mathbb{C}^{m\times n}_{\neq 0} \to \mathbb{R}^{m\times m}_{\geq 0}$ be the mapping defined by

$$f_{mn}(A) = \mathrm{diag}(w(A))^{-1}, \tag{3.8}$$

where w is defined similarly as in Definition 3.1.

The above mapping is well defined as $w(A)_i \neq 0$ for all i, due to the fact that A has no zero row or column. We note that we can apply the mapping f_{nm} to the adjoint of A, as $A^* \in \mathbb{C}^{n\times m}_{\neq 0}$. It is easily seen that $f_{nm}(A^*) = f_{nm}(A^t)$.

Definition 3.7. Let $g : \mathbb{C}^{m\times n}_{\neq 0} \to \mathbb{C}^{m\times n}_{\neq 0}$ be the mapping defined by

$$g(A) = f_{nm}(A^*)A^*f_{mn}(A), \tag{3.9}$$

for all $A \in \mathbb{C}^{m\times n}_{\neq 0}$.

Lemma 3.8. *If $A \in \mathbb{C}^{m\times n}_{\neq 0}$, then $\lambda(g(A)A) \in [0,1]$. If furthermore A is of rank m, then $\lambda(g(A)A|_{\mathrm{Im}(g(A)A)}) \in (0,1]$. Moreover* $\mathrm{Ker}(A) = \mathrm{Ker}(g(A)A)$.

Proof. We claim that $A^*f_{mn}(A)A$ is positive semi-definite. This follows easily, since $f_{mn}(A)$ is a positive definite diagonal matrix, and $z^*A^*f_{mn}(A)Az = \left\|f_{mn}(A)^{1/2}Az\right\| \geq 0$ for any $z \in \mathbb{C}^n$. Now $g(A)A$ is the product of a positive definite matrix $f_{nm}(A^*)$ and a semi-positive definite matrix $A^*f_{mn}(A)A$, therefore its eigenvalues are positive. Much like in the proof of lemma 3.2, we use Gershgörin-Hadamard and Hölder's inequality to show that :

$$\begin{aligned} \lambda_{max}(g(A)A) &\leq max|g(A)A|\,u \\ &\leq \|Af_{nm}(A^*)\|_1\,\|f_{mn}(A)A\|_\infty \\ &= \max f_{nm}(A^*)\,|A^*|\,f_{mn}(A)\,|A|\,u. \end{aligned}$$

We remark that $f_{nm}(A^*)\,|A^*|\,f_{mn}(A)\,|A|$ is a stochastic matrix and therefore the righthand-side of the above inequality is smaller than 1.

We claim that $\mathbb{C}^n = \mathrm{Im}(g(A)A) \oplus \mathrm{Ker}(g(A)A)$. If this is the case, $\mathrm{Ker}(g(A)A|_{\mathrm{Im}(g(A)A)}) = \{0\}$, as we restrict $g(A)A$ to the complement of its kernel. Since we have already shown that $\lambda(g(A)A) \geq 0$, and $g(A)A|_{\mathrm{Im}(g(A)A)}$ has no zero eigenvalue, the result follows. It remains to show the claim. Since A is of rank m, the orthogonal complement of its range is a space of dimension $n-m$, therefore the nullspace of A^* also has dimension $n-m$, it follows that the rank of A^* is m. We now remark that $\mathrm{Im}(f_{mn}(A)A) = \mathbb{C}^m$, therefore due to the construction of $g(A)A$, it has equally to be of rank m. It is easily seen that $\mathrm{Im}(g(A)A) = \mathrm{Im}(A^*)$, therefore $\mathrm{Im}(g(A)A)^\perp = \mathrm{Ker}(A)$. Since $\mathrm{Ker}(A) \subseteq \mathrm{Ker}(g(A)A)$ and both spaces are closed and of dimension $n-m$, they must be equal and the claim follows. □

The following theorem can be proven in a similar way as we did for Lemma 3.3.

Theorem 3.9. *Let $A \in \mathbb{C}^{m\times n}_{\neq 0}$ be of rank m. The iteration matrix $Q = I - \alpha g(A)A|_{\mathrm{Im}(g(A)(A)}$ is convergent for any $\alpha \in (0,2)$.*

$\mathrm{Ker}(AR)$ and $\mathrm{Im}(RA)$ determine the nature of the inverse, the proof is identical to that of Lemma 3.5.

Lemma 3.10. *For any matrix A, the generalized inverse, called G generated by matrix $R = g(A)$, is reflexive, $f_{nm}(A^*)^{-1}$-minimal-norm and $f_{mn}(A)$-least squares.*

3.3. Cimmino

We can further generalize our construction by choosing the l_p norm in definition 3.6 instead of the l_1 norm and replace equation (3.9) by:

$$g(A) = f_{nm}(A^*)^k A^* f_{mn}(A)^l,$$

where $k, l \in \mathbb{R}_+$ and $k + l = 2$.

The classical gain matrix of Cimmino can now be obtained by choosing $\alpha = \frac{2}{n}$ in Theorem 3.9, $k = 0, l = 2$ and the l_2 norm. Our matrix choice corresponds to $k = l = 1$, $\alpha = 1$ and the l_1 norm.

3.4. Preconditioning a system

The Gauss-Seidel scheme can be seen as (1.1) preconditioned by matrix $(I - RL)^{-1}R$, with R chosen to be equal to the Jacobi matrix J:

$$(I - RL)^{-1}RAx = (I - RL)^{-1}Rb,$$

where L is defined as follows : $L_{ij} = -a_{ij}$ if $i > j$ and 0 otherwise. We can use this scheme with a more general choice of the matrix R, such as $g(A)$. We know that this scheme converges when, $\rho(I - RA|_{\mathrm{Im}(I-RA)}) < 1$, which is the case for our matrix $g(A)$ (see Theorem 3.9).

4. Some solution schemes

In this paragraph we will use our preconditioning matrix in several classical algorithms and show their convergence. Throughout this paragraph, R will be equal to $\hat{f}(A)$, whenever $A \in \mathbb{C}_+^{m\times m}$ and $g(A)$, whenever $A \in \mathbb{C}_{\neq 0}^{m\times n}$. We will consider the following system :

$$RAx = Rb. \tag{4.10}$$

The next proposition is a direct consequence of Lemma 3.8, Theorem 3.9 and Lemma 3.3.

Proposition 4.1. *Let $A \in \mathbb{C}_{\neq 0}^{m\times n}$ be of rank m or $A \in \mathbb{C}_+^{m\times m}$ and let R be the corresponding gain matrix. Then* $\mathrm{Ker}(RA) = \mathrm{Ker}(A)$ *and* $\rho(I - RA|_{\mathrm{Im}(I-RA)}) < 1$.

4.1. Richardson-Tanabe

The Richardson-Tanabe scheme for (4.10) is given by

$$x_0 \in \mathbb{C}^n \qquad x_{k+1} = x_k + \alpha R(b - Ax_k),$$

where $\alpha \in (0, \frac{2}{\rho(RA)})$. We know that this scheme converges when $\mathrm{Ker}(RA) = \mathrm{Ker}(A)$, (1.1) has a solution, and $\rho(I - RA|_{\mathrm{Im}(I-RA)}) < 1$ (see [19]). Due to Proposition 4.1, we see that for our choice of R, this scheme converges. We choose $\alpha = 1$ for the sake of simplicity.

4.2. Schultz-Hotelling-Bodewig

In this algorithm we calculate successive approximations of the inverse (or generalized inverse) of A.

$$\begin{aligned} f_0(A) &= R \\ f_k(A) &= f_{k-1}(A)[2I - Af_{k-1}(A)] \\ &= [2I - f_{k-1}(A)A]f_{k-1}(A). \end{aligned}$$

Let $C^j_{2^k} = \frac{2^k!}{j!(2^k-j)!}$. By simple induction we see that we can write

$$\begin{aligned} f_k(A) &= \sum_{j=1}^{2^k}(-1)^{j-1}C^j_{2^k}(RA)^{j-1}R \\ f_k(A) &= R\sum_{j=1}^{2^k}(-1)^{j-1}C^j_{2^k}(AR)^{j-1}. \end{aligned} \tag{4.11}$$

Theorem 4.2. *Let $A \in \mathbb{C}^{m\times n}_{\neq 0}$ or $A \in \mathbb{C}^{m\times m}_{+}$ be of rank m and let R be the corresponding gain matrix. Then $\lim_{k\to\infty} f_k(A) = A^-$.*

Proof. We remark first of all that $RA = RA(RA)^-RA$, so we can write

$$\begin{aligned} f_k(A)A &= \sum_{j=1}^{2^k}(-1)^{j-1}C^j_{2^k}(RA)^{j-1}RA \\ &= \sum_{j=1}^{2^k}(-1)^{j-1}C^j_{2^k}(RA)^{j-1}RA(RA)^-RA. \end{aligned}$$

It is easily seen that

$$\sum_{j=1}^{2^k}(-1)^jC^j_{2^k}(RA)^j = (I-RA)^{2^k} - I.$$

We remark that for any matrix X, we have $\|X^k\| \leq \|X\|^k = \mu_{max}(X)^k \leq \rho(X)^k$. Due to Proposition 4.1 and the above we see that $(I-RA)^{2^k}$ tends to zero. Therefore

$$\begin{aligned} \lim_{k\to\infty} f_k(A)A &= \lim_{k\to\infty}[I-(I-RA)^{2^k}](RA)^-R \\ &= (RA)^-RA. \end{aligned}$$

We remark that since A is of rank m its range is $\mathbb{C}^m$, therefore $\lim_{k\to\infty} f_k(A) = (RA)^-R$. It remains to be shown that $(RA)^-R = A^-$. We know that $RA(RA)^-RA = RA$ and $\mathrm{Ker}(A) = \mathrm{Ker}(RA)$, it follows that $A(RA)^-RA = A$. Due to the definition of $(RA)^-$ we trivially have $(RA)^-RA(RA)^-R = (RA)^-R$. We have shown that $(RA)^-R = A^-$. □

It is easily seen from equation (4.11) that if A is semi-positive, then $f_k(A)A$ is a stochastic matrix and therefore the iteration matrix $Q_k = I - f_k(A)A$ has eigenvalue zero with associated eigenvector u.

4.3. Conjugated Gradient

In this section we will describe the conjugated gradient algorithm applied to our preconditioned system. In order to do so we will consider preconditioning the following system

$$A^* f_{mn}(A)Ax = A^* f_{mn}(A)b,$$

with $f_{nm}(A^*)$.

Theorem 4.3. *Let $A \in \mathbb{C}^{m\times n}_{\neq 0}$. Then $A^* f_{mn}(A)A$ is self-adjoint and positive-semi definite. Moreover $A^* f_{mn}(A)A$ is positive definite if and only if A is of rank n.*

Proof. It is easily seen that $A^* f_{mn}(A)A$ is self-adjoint. Furthermore we remark that

$$z^* A^* f_{mn}(A)Az = \left\| f_{mn}(A)^{\frac{1}{2}} Az \right\| \geq 0$$

and that $z^* A^* f_{mn}(A)Az > 0$ if and only if $Az \neq 0$. We note finally that $\mathrm{Ker}(A) = \{0\}$ if and only if A is of rank n. The theorem follows. □

We remark that depending on the dimension of A, $\mathrm{rank}(A) = n$, might be impossible. We obtain the following algorithm, that converges if $\mathrm{rank}(A) = n$ [7] :

$$\begin{array}{rcc}
x_0 & \in & \mathbb{C}^n \\
p_0 & = & f_{nm}(A^*)^{-1} r_0 \\
r_0 & = & A^* f_{mn}(A)b - A^* f_{mn}(A)Ax_0 \\
z_0 & = & p_0 \\
\alpha_k & = & \frac{\langle r_k, z_k\rangle}{\langle A^* f_{mn}(A)Ap_k, p_k\rangle} \\
x_{k+1} & = & x_k + \alpha_k p_k \\
r_{k+1} & = & r_k - \alpha_k A^* f_{mn}(A)Ap_k \\
z_{k+1} & = & f_{nm}^{-1} r_{k+1} \\
\beta_{k+1} & = & \frac{\langle r_{k+1}, z_{k+1}\rangle}{\langle r_k, z_k\rangle} \\
p_{k+1} & = & z_{k+1} + \beta_{k+1} p_k
\end{array}$$

4.4. Kaczmarz

The usual form of this algorithm [2] is the following

$$x_j^{k+1} = x_j^k + \alpha_k \frac{a_{i_k,j}}{\|a_{i_k}\|_2^2}(b_{i_k} - \langle a_{i_k}, x^k\rangle), \tag{4.12}$$

where $\alpha_k \in (0, 2)$, $i_k = k \mod (m) + 1$ and a_{i_k} is the i_kth row-vector of A.

We propose to generalize the above algorithm by using the l_p norm instead of the l_2 norm. The modified algorithm then becomes :

$$x_j^{k+1} = x_j^k + \alpha_k \frac{a_{i_k,j}}{\|a_{i_k}\|_p^2}(b_{i_k} - \langle a_{i_k}, x^k\rangle). \tag{4.13}$$

Theorem 4.4. *The l_p generalized Kaczmarz algorithm converges.*

Proof. We know that the Kaczmarz algorithm converges for $\alpha_k \in (0,2)$. Furthermore any two norms on the finite dimensional space $\mathbb{C}^n$ are equivalent. Therefore we can find some scalar $\lambda_p > 0$ such that

$$\frac{\alpha_k}{\|x\|_p^2} \leq \frac{\alpha_k}{\lambda_p^2 \|x\|_2^2}.$$

Choosing α_k such that $\frac{\alpha_k}{\lambda_p^2} \in (0,2)$, therefore assures convergence. □

5. Numerical resolution

In this paragraph we will test the solution schemes defined in Section 4 in practice. We will test the following 16 methods (see table (1)) on our tests cases. The first 10 are well-known methods, and the methods 11 to 16 are those proposed in this paper.

We will use the notation *RT* for Richardson-Tanabe, SHB for Schultz-Hotelling-Bodewig and CG for conjugated gradients.

1	Gauss	GAUSS	10	Cimmino, CG scheme	CIMGC
2	Pseudo-Inverse	PINV	11	NAM, generalized Kaczmarz scheme with matrix NA^*M	NAMKAC
3	CG	GC	12	N, generalized Kaczmarz scheme with matrix N	MKAC
4	CG of minimal residue	RESMIN	13	NAM, RT scheme, with relaxation factor α	αNAMRT
5	Kaczmarz	KACZ	14	NAM, SHB scheme	NAMSHB
6	Kaczmarz, RT scheme with matrix K	KACZRT	15	M, SHB scheme	MSHB
7	Kaczmarz, SHB scheme	KACSHB	16	N, pre-conditioned CG scheme	NGC
8	Cimmino, RRT scheme with matrix C	CIMRT	17	NAM, pre-conditioned CG scheme	NAMGC
9	Cimmino, SHB scheme	CIMSHB			

Table 1. List of the different schemes tested.

For each family of algorithms we will only show that algorithm that gives the best results, i.e., those that not only obtain a small residue but also an accurate estimate of the solution. We sometimes comment on some algorithms even though they do not appear on the figures.

Finally we briefly mention an application of the proposed algorithms for the solution of a large scale system in Quantum mechanics.

5.1. Nagasaka-Tanabe Matrix

The following test has been proposed by Nagasaka [14] in order to study the propagation of numerical errors in the Gauss algorithm and was later reused by Tanabe [18]. In this test the matrix A is the following tridiagonal positive definite 84×84 matrix of condition-number $\kappa(A) = 5.73 \cdot 10^{16}$:

$$A = \begin{bmatrix} 6 & 1 & & & & 0 \\ 8 & 6 & 1 & & & \\ & . & . & . & & \\ & & & 8 & 6 & 1 \\ 0 & & & & 8 & 6 \end{bmatrix} \tag{5.14}$$

As we know that our algorithm will converge in a single iteration when $x = u$, we will not show the results of this test. Instead we will consider the case, where $x_i = i$. The vector b in equation (1.1) can (of course) be obtained easily. Figure (1) shows the convergence speed of the most performing algorithms on this test case.

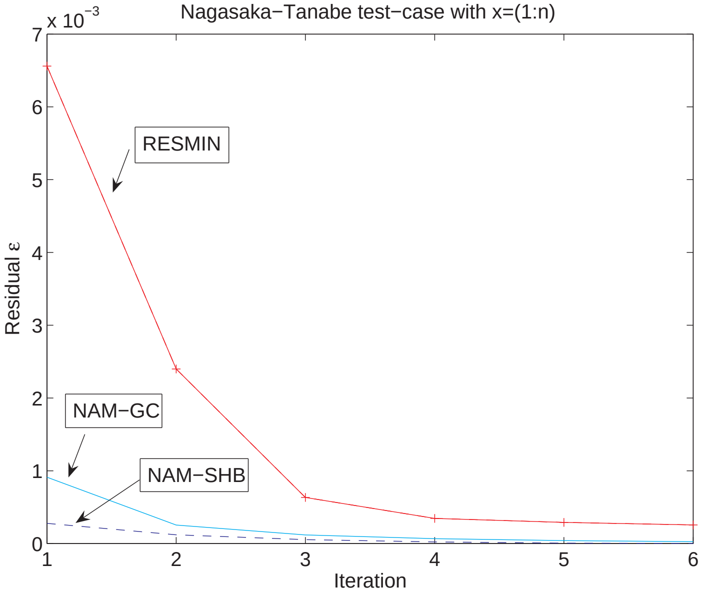

FIGURE 1. Convergence Speed of some tested Algorithms in the Nagasaka case

As is well known Gauss' algorithm is numerically unstable for such matrices. The *RESMIN* algorithm is however unable to reasonably approximate the solution x better than 10^{-5}. We have observed that the family of Kaczmarz algorithms (*KACZ*, *KACZRT*, *KACSHB*) has difficulties to properly estimate the last component of x, despite of the small error ε. The *NA*M* family of algorithms has a small error ε and reasonably estimates the last components of x. For the *NA*MRT* algorithms, a good choice of the relaxation factor α allows us to reduce the number of iterations by a factor 2.

5.2. Hilbert Matrix

We will test our algorithms with the Hilbert matrix ($A_{ij} = \frac{1}{i+j-1}$), as we wish to determine their behaviour in the presence of a bad condition number. We will therefore test our algorithm with the 15 × 15 Hilbert matrix (We recall that its condition number is $8.48 \cdot 10^{17}$). Again, as we know, due to Lemma 2.3, our algorithm will converge in a single iteration when $x = u$, we will not show the results of this test-case. Instead we will consider solving the system when $x_i = i$, as before.

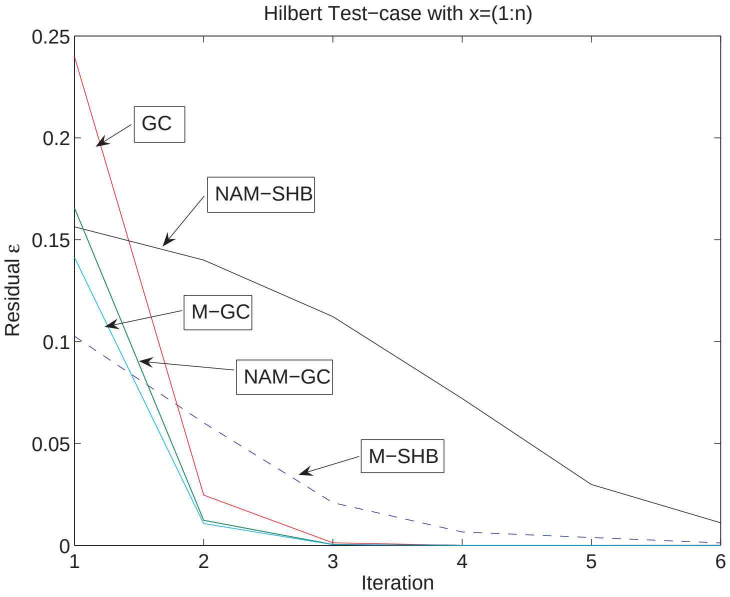

Figure 2. Convergence Speed of some tested Algorithms in the Hilbert case

On figure (2) we observe the excellent behaviour of the conjugated gradient schemes, specifically adapted to this type of matrix. The *NAMGC* and *MSHB* however give solutions of comparable quality in only half of the number of iterations. We have finally observed a very slow convergence (speed) of the Richardson-Tanabe schemes (*KACZRT*, α*NAMRT*). Finally, *MSHB* converges faster than *NAMSHB*, but give a solution of comparable quality in the end.

5.3. Lower diagonal dominant tri-diagonal Matrix

Our last test-case involves a tri-diagonal matrix of dimension 84, where the lower diagonal is dominant. Our matrix A is therefore the following matrix :

$$A = \begin{bmatrix} 1 & 1000 & & & 0 \\ 10000 & 1 & . & & \\ & . & . & . & \\ & & . & 1 & 1000 \\ 0 & & & 10000 & 1 \end{bmatrix} \tag{5.15}$$

We will use again the vector x, where its ith component is i. The whole difficulty of the test-case is estimating the last component of x. For the Gauss method the above matrix appears to be singular, therefore we have not been able to use this method. Figure (3) again shows the convergence speed of some of the algorithms.

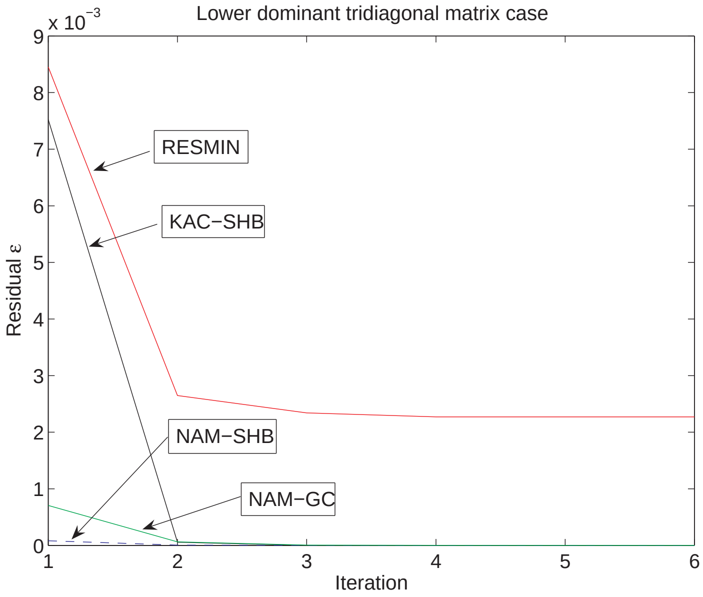

Figure 3. Convergence Speed of some tested Algorithms in the tri-diagonal case

Except for the *NA*M* schemes, only *RESMIN* allows us to reasonably estimate x_n, with only an approximation error of 8.6% (All other algorithms have a 90% error). However this algorithm offers only little precision ε. Except for *MKAC*, the *NA*M* algorithms give a solution with less than 2% approximation error for the last component of x. It is clear that if we had used an upper-dominant matrix, the "edge"-effect would have been on the first component of x, with similar results.

5.4. **Large Scale Systems**

We prospect to test the proposed algorithms on several large scale systems. This section briefly mentions what three applications are thought about.

5.4.1. *Three dimensional Schrödinger Equation*

In [15, 16], the authors tackle the optical properties of self-assembled III-V and IV-IV nanostructures, in particular for quantum information and telecommunication devices. The aim is to solve the strain-dependent Schrödinger equation in 8 band k.p theory.

A very efficient method is proposed in the 3D numerical code JAVEL. For a very simplified linear system of dimension 343, applying the schemes *NAMSHB* and *NAMGC*, we obtain the following results (figure (4)) which are encouraging.

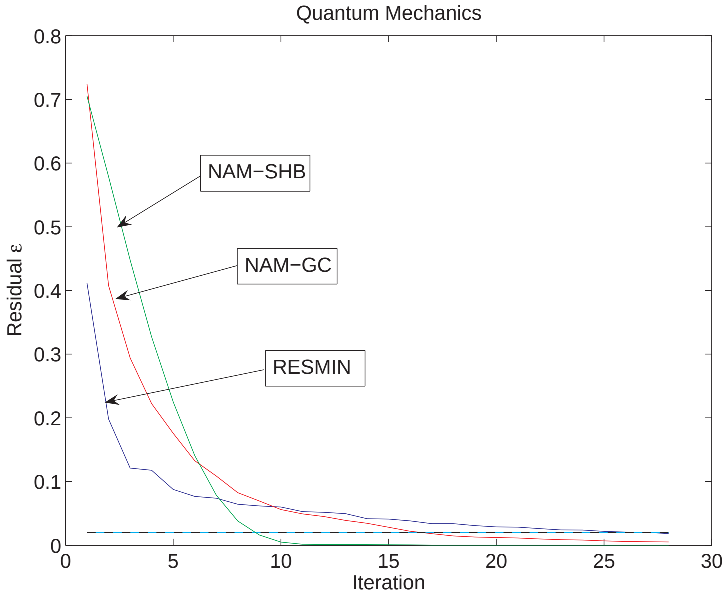

Figure 4. First results on a small Quantum mechanics test case

The horizontal line corresponds to the minimal requested accuracy. As we can observe, the *NAMSHB* algorithms allows us to obtain required accuracy, in less iterations than *RESMIN*. The former using matrix products, the latter only matrix-vector products. Work is in progress in order to formulate proposed numerical schemes using only products of matrices and vectors.

5.4.2. *Unit Commitment*

The Unit Commitment Problem (UCP) consists of defining the minimal-cost power generation schedule for a given set of power plants, thermal and hydraulic. We will focus on hydraulic valley management. The interconnected reservoirs and productions units, once modeled give way to a large linear system (from 5000 to 15000 variables).

We use Lagrangian relaxation, price decomposition and other optimization techniques to solve the UCP [6,13], exploiting its highly decomposable structure. In this context, the sub-problem related to an hydraulic valley can be stated as a linear program.

The interior point method is currently used for solving this linear program (which is itself solved at each iteration of the price decomposition algorithm, typically 400 times). As is known, the linear system solved at each iteration of the interior point algorithm becomes more and more badly conditioned. The proposed algorithms might therefore improve the solution quality.

5.4.3. *Electromagnetics*

In electromagnetic scattering problems, the scattered fields of interest (electric and/or magnetic depending on the applications) can be expressed using some linear integro-differential equations linking physical parameters of the configuration to the scattered field through some dyadic Greens functions. The numerical discretization of the equations (using a method of moments) leads to a linear system to be solved. If in most of the cases, the system is well conditioned and can be easily solved, in some particular applications (non destructive evaluation using eddy-current or geophysical exploration) the convergence of the classical iterative algorithms is far to be optimal (large number of iterations, slow convergence). The algorithms developed in this paper could be tested in such configurations and might be a way to greatly improve the speed/quality of some electromagnetic codes.

6. Conclusions

Estimating efficiently the solution of an ill-conditioned linear system of equations is tricky.

The new approach we have developed, based on properties of stochastic matrices and l_p norms, allows us to define in two steps algorithms for efficiently solving any ill-conditioned, and possibly singular, linear system of equations. First, we construct an approximation of a generalized inverse of the matrix of the system (which can be complex, rectangular and without specific structure). The second step consists in using this matrix in classical schemes as Richardson-Tanabe, Schultz-Hotelling-Bodewig or conjugate gradients. We also propose a generalization of the Kaczmarzs scheme, initially defined for l_2 norm, as other possible scheme of resolution.

Even considering algorithms obtained with l_1 norm, we reported outstanding performances (performance is understood as the capacity to retrieve as best as possible the exact solution). In this way, for any matrix A or more specifically for positive semi-definite hermitian matrix, at least one of the proposed algorithms is more efficient than the algorithms we've tested. Hence :

- when A is regular, hermitian, positive-definite, we suggest to use the matrix $R = \hat{f}(A)$, in the preconditioned conjugate gradients scheme since, regarding the ill-conditioning snag, this scheme is more efficient and more robust than the classical conjugate gradients;
- for any matrix A, using the matrix $R = \alpha f_{nm}(A^*) A^* f_{mn}(A)$ in the Schultz-Hotelling-Bodewig's scheme or in the conjugate gradients scheme gives very robust results regarding the ill-conditioning and error propagation.

For some ill-conditioned matrices with specific structure (tridiagonale, upper or lower dominance, . . .), the proposed algorithms are able to correctly estimate some components of the solution-vector, whereas none of the known algorithms tested is able to do so.

When an existing algorithm is more efficient than those proposed (it is the case for the conjugate gradients with the Hilbert matrix test-case), one of proposed algorithms gives a solution with equivalent quality in approximatively two times less iterations.

Finally, regarding the first tests we have reported, it seems that the proposed algorithms are at least as efficient as the classical techniques tested. On our test-cases, they even seem to be more efficient. In addition, these algorithms are universal since they don't need any restrictive hypothesis or conditions for application.

One can ask, when facing solving a pathological (huge, very ill-conditioned, specific structure, ...) linear system of equations, it would be possible to define low-cost procedures for extracting useful information of the solution without solving the system. For example the extreme values of the solution, the mean, the variance, etc. The promising approach of stochastic matrices developed here seems to be an interesting tool for exploring this way.

The authors would like to thank Sebastien Sauvage for providing us with data related to the Schrödinger equation application

References

[1] A. Ben-Israël and T.N.E. Greville. *Generalized inverses : Theory and applications.* Wiley-Interscience, 1974.

[2] A. Björck and T. Elfving. Accelerated projection methods for computing pseudo-inverse solutions of a system of linear equations. *BIT*, 19:145–163, 1979.

[3] E. Bodewig. *Matrix Calculus*. Interscience Ed., New York, 2 edition, 1959.

[4] Y. Censor. Row-action methods for huge and sparse systems and their applications. *SIAM review*, 23(4), 1981.

[5] G. Cimmino. Calcolo approssimato per le soluzioni dei sistemi dei equazioni lineari. *Ricerca Sci. Progr. Tecn. Econom. Naz.*, 9:326–333, 1938.

[6] G. Cohen and D.L. Zhu. Decomposition-coordination methods in large-scale optimization problems. the non-differentiable case and the use of augmented langrangiens. 1, 1983.

[7] M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952.

[8] R.A. Horn and C.R. Johnson. *Matrix analysis*. Cambridge University Press, 1999.

[9] H. Hotelling. Further points on matrix calculations and simultaneous equations. *Ann. Math. Statist.*, 14:440–441, 1943.

[10] H. Hotelling. Some new methods in matrix calculation. *Ann. Math. Statist.*, 14:1–34, 1943.

[11] S. Kaczmarz. Auflösung von systemen linearer gleichlungen. *Bulletin de l'Académie Polonaise des Sciences et Lettres*, A:355–357, 1937.

[12] P. Lascaux and R. Theodor. *Analyse numérique matricielle appliquée à l'art de l'ingénieur*. Masson Editeur, Paris, 1987.

[13] C. Lemaréchal and C. Sagastiźabal. An approach to variable metric bundle methods. *Lecture notes in Control and Information Science*, (197):144–162, 1994.

[14] H. Nagasaka. Error propagation in the solution of tridiagonal linear equations. *Information Processing in Japan*, 5:38–44, 1965.

[15] S. Sauvage, P. Boucaud, F. Bras, G. Fishman, R.P.S.M. Lobo, F. Glotin, J.-M. Ortéga, and J.-M. Gérard. Long polaron lifetime in inas/gaas self-assembled quantum dots. *Phys. Rev. Lett*, (88):177–402, 2002.

[16] S. Sauvage, P. Boucaud, F. Bras, G. Fishman, R.P.S.M. Lobo, F. Glotin, R. Prazeres, J.-M. Ortéga, and J.-M. Gérard. Polaron relaxation in inas/gaas self-assembled quantum dots. *phys. stat. sol. (b)*, 238(2):254–257, 2003.

[17] G. Schultz. Iterative berechnung der reziproken matrix. *Z. angew. Math. Mech.*, 13:57–59, 1933.

[18] K. Tanabe. Projection method for solving a singular system of linear equations and its applications. *Numerische Mathematik*, 17:203–214, 1971.

[19] K. Tanabe. Characterization of linear stationary iterative processes for solving a singular system of linear equations. *Numerische Mathematik*, 22:349–359, 1974.