Title: **REDUCTION OF THE VARIANCE IN MONTE CARLO ALGORITHMS FOR SOLVING SYSTEMS OF LINEAR ALGEBRAIC EQUATIONS**

Scientific advisor: **Todor Dimov**

# THESIS SUMMARY

## 1 General description of the thesis

The presented thesis is devoted to the minimization of the probable error in calculating linear functional of the solution of a **s**ystem of **l**inear **a**lgebraic **e**quations (SLAE).

Monte Carlo methods are a powerful tool for solving different problems in the fields of mathematics, phisics and engineering. It is known that they provide statistical estimations for a functional of the solution using sample of a certain random variable whose mathematical expectation is equal to the given functional. These methods are used when not very accurate solution is needed (in the real-life computations the required accuracy is about 1-5 %). The MCM is very useful when one is interested in finding inverse matrix, because the problem of estimating the inverse matrix can be present as a problem of solving SLAE (see [5]).

There are several basic advantages of these algorithms. It is well known that Monte Carlo algorithms are parallel algorithms. They have high parallel efficiency when parallel computers are used. Monte Carlo algorithms are also very efficient when the problem under consideration is too large or too intricate for other treatment. One of the most important advantages of these algorithms is that they can be used for evaluating only one component of the solution or some linear form of the solution. In this case it is not necessary to perform the all computational work which is needed for obtaining the complete solution.

In general, there are two classes of Monte Carlo numerical algorithms - direct algorithms and iterative algorithms. The direct algorithms obtain the approximate solution of a problem in a finite number of steps, and contain only stochastic error. The iterative Monte Carlo methods use an approximation of the solution obtaining a certain number of significant digits. The iterative Monte Carlo methods have two types of error - stochastic and systematic. The systematic error depends on the number of performed iterations of the used iterative method, whereas the stochastic error depends on the probabilistic nature of the method.

It is well known [1] that iterative Monte Carlo methods are preferable for solving large sparse systems (succh as those arising from approximations of partial defferential equations). Such methods are good for diagonally dominant systems for which the rate of the convergence is high.

**Aim of the master thesis**:
Research, modification and numerical testing of known algorithms of type Monte Carlo for calculating linear functional of the solution of a system of linear algebraic equations.
According to this aim **the main problems** of the master thesis are:

1. To describe and study a method with zero variance using information for solution of the conjugate problem.

2. To study efficiency of a Monte Carlo algorithm for solving SLAE for two concrete choices of $p_{ij}$ (transition probabilities). To propose a modification when the used information is connected only with the elements of the given system.

3. To make numerical tests with the modified Monte Carlo algorithm for solving SLAE with large general sparse matrices (with optimal memory loading).

**Thesis structure**: The thesis consists of Introduction, four Chapters, and List of references. The text contains 60 pages and includes 5 tables.
**Chapter 1**: A short description of Monte Carlo (direct and iterative) methods and discrete Markov chain are given. The considered problem is described.
**Chapter 2**: The theorem of Ermakov and Mikhailov [6] dor solving integral equations is applied dor solving SLAE. The efficiency of the constructed algorithm is studied.
**Chapter 3**: An algorithm, given in [10], with transition probabilities proportional to the elements of the iterative matrix ([8]) is the base for solving the second main problem. A modification of this algorithm with balancing the iterative matrix or the right hand side of the system is proposed. In result of this procedure the variance decreases. The balancing is realized using the different relaxation paarmeters.
**Chapter 4**: An algorithm for general sparse matrices is described. The calculations are performed only with nonzero elements. This gives solution of the third main problem for increasing the computational efficiency of the considered algorithm.

# 2 Problem statement

Consider the following SLAE:

$$Ax = b, \qquad A \in \mathbb{R}^{m\times m}, \quad x, b \in \mathbb{R}^m. \tag{1}$$

This system can be presented in the following form (if only $a_{ii} \neq 0, \forall i = 1, \ldots, m$) using the Jacobi relaxation iterative method with relaxation parameter $\gamma \in (0; 1]$:

$$x = Lx + f, \tag{2}$$

where

$$L = (l_{ij})_{i,j=1}^{m} \in \mathbb{R}^{m\times m}, \qquad f = (f_1, \ldots, f_m)^T$$

and

$$l_{ij} = \begin{cases} 1-\gamma, & i = j \\ -\gamma\dfrac{a_{ij}}{a_{ii}}, & i \neq j \end{cases} \qquad i,j = 1,\ldots,m$$

$$f_i = \gamma\frac{b_i}{a_{ii}}, \qquad i = 1,\ldots,m.$$

We are interested in the calculating linear functional of the solution:

$$(x,h) = \sum_{i=1}^{m} x_i h_i, \tag{3}$$

where $h \in \mathbb{R}^m$ is a given vector.

To find one component of the solution, for example the $i_0$-th component of $x$, we choose $h = e(i_0) = (0,\ldots,0,1,0,\ldots,0)^T$, where the one is in the $i_0$-th place. To calculate the functional $(x,h)$, an iterative Monte Carlo method is used. We construct a random variable $X$, whose mathematical expectation is equal to the linear form (3), using discrete Markov processes with a finite set of states (finite discrete Markov chains). Then the computational problem becomes one of calculating repeated realizations of $X$ and of combining them into an appropriate statistical estimator of $(x,h)$.

*The problem of the minimization of the probable error, which is equivalent to the minimization of the variance of the constructed random variable is studied.*

# 3 Chapter two. Minimization of the variance using a priori information about the solution of the conjugate problem

In this chapter a theorem for minimization of the variance is proved. The result is achieved through an appropriate choice of the initial and transition probabilities. This theorem is analogous to the theorem of Ermakov and Mikhailov [6] for integral operators.

**Definition 3.1** *Conjugate problem of (2):*

$$x^* = L^T x^* + h \qquad (L^T = (l^*_{ij})_{i,j=1}^{m}).$$

**Theorem 3.1** *Let the matrix $A$ is diagonally dominant and*

$$\pi_i = \frac{f_i x^*_i}{(f,x^*)},\ p_{ij} = \frac{l^*_{ij}x^*_j}{(L^T x^*)_i},\ \forall i,j = 1,\ldots,m.$$

*Then* $\mathrm{E}X = (x, h)$ *and* $\mathrm{D}X = 0$, *where*

$$X = \sum_{i=0}^{\infty} W_i h_{s_i}.$$

The theorem is proved using the following statement:

**Lemma 3.1** *Let the random variable is given:*

$$Y^{(0)} = W_0 h_{s_0} \qquad Y^{(k)} = \sum_{i=0}^{k-1} W_i h_{s_i} + W_k x^*_{s_k},\ k \in \mathbf{N}\backslash\{0\}.$$

*Then*

$$\mathrm{E}Y^{(k)} = (x, h) \qquad \mathrm{D}Y^{(k)} = 0,\ k \in \mathbf{N}.$$

The statement, which is proved in Theorem 3.1, has a theoretical meaning, but its practical application is connected with some obvious disadvantages. The solution $x^*$ is unknown. On the other hand, it necessary to construct an infinite Markov chain to obtain zero variance. Obviously, this is impossible.

# 4 Chapter three. Minimization of the variance using a priori information about the solution of the given system

## 4.1 Theoretical background

A theorem, which is given in [4], is the base of considered algorithm.

**Theorem 4.1** *Let the matrix* $A$ *is diagonally dominant. Then* $\mathrm{E}X = (x, h)$, *where*

$$X = \frac{h_{s_0}}{\pi_{s_0}} \sum_{i=0}^{\infty} W_i f_{s_i}$$

$$\left( W_0 = 1, \qquad W_i = W_{i-1} \frac{l_{s_{i-1}s_i}}{p_{s_{i-1}s_i}},\ i \geq 1 \right).$$

Obviously, if $A$ is a diagonally dominant matrix, then the elements of the matrix $L$ must satisfy the following condition:

$$\sum_{j=1}^{m} |l_{ij}| < 1,\ \forall i = 1, \ldots, m. \tag{4}$$

It is well known that property (4) is a sufficient condition for convergence of the Neumann series, i.e.

$$x = \lim_{k\to\infty} x^{(k)}, \qquad x^{(k)} = \sum_{i=0}^{k-1} L^i f + L^k x^{(0)},\ k > 0.$$

It is clear that every iterative algorithm uses a finite number of iterations $k$. In the algorithm, following finite sum

$$X^{(k)} \doteq \frac{h_{s_0}}{\pi_{s_0}} \sum_{i=0}^{k} W_i f_{s_i},\ (k \in \mathbf{N}), \qquad \mathrm{E}X^{(k)} = (x^{(k)}, h)$$

is computed, where $x^{(k)}$ is the $k$-th iterative approxiamtion of the solution $x$ (using $x^{(0)} \doteq f$). The trucation parameter $k$ is obtained from the condition that the difference between the stochastic approximation of two successive approximations is smaller than a given sufficiently small parameter $\varepsilon$.

## 4.2 Advantages and disadvantages of the algorithm with respect to the two choices of the transition probabilities

The transition probabilities are chosen to minimize the variance of the random variable. The two ideas are:

1. $$p_{ij} = \frac{1}{n_i}, \qquad i, j = 1, \ldots, m,$$

   where $n_i$ is the number of nonzero elements in the $i$-th row of $L$;

2. $$p_{ij} = \frac{|l_{ij}|}{\sum_{j=1}^{m} |l_{ij}|} \qquad i = 1, \ldots, m. \tag{5}$$

The algorithm with transition probabilities which are proportional to the $|l_{ij}|$ is called Monte Carlo almost optimal algorithm (see [3] and [8]). Let us note that the second algorithm has the following imporatant property:

$$\frac{l_{ij}}{p_{ij}} = const(i) \quad \text{for all} \quad i = 1, \ldots, m,$$

which solves the problem with minimization of the variance to some extent. This algorithm can become optimal when additional conditions for the system are given.

The following statement is valid:

**Statement 4.1** *The system (1) (corresponding system (2)) is given. Let the matrix $A$ is diagonally dominant, the element of $L$ have the same sign, and transition probabilities are given with the formula (5). Suppose that*

$$l \doteq l_i = l_j \qquad \forall i \neq j,\ i,j \in \{1,\ldots,m\},$$
$$f \doteq f_i = f_j$$

*where*

$$l_i = \sum_{j=1}^{m} |l_{ij}|, \qquad i = 1,\ldots,m.$$

*Then*

$$\mathrm{D}X = 0, \qquad X = \frac{h_{s_0}}{\pi_{s_0}} \sum_{i=0}^{\infty} W_i f_{s_i}.$$

**Remark 4.1** *It is true that* $\mathrm{D}X^{(k)} = 0, \forall k \in \mathbf{N}$ *besides* $\mathrm{D}X = 0$. *It is very important that the conditions of the statement (4.1) guarantee a zero variance without realization of a boundary transition (difference with Theorem 3.1).*

For every row of matrix $A$ (if $|a_{ii}| \neq 0$) a parameter $k_i$ is defined:

$$k_i \doteq \frac{\sum\limits_{j=1, j\neq i}^{m} |a_{ij}|}{|a_{ii}|}, \qquad i = 1,\ldots,m.$$

The proved statement gives the conditions for obtaining a zero variance - the row balancing the iterative matrix and balancing the right hand side of the system.

## 4.3 Using different relaxation parameters

- Balancing the iterative matrix:

  We use the notation:

  $$k_{i_{max}} \doteq \max_{1\leq i\leq m} k_i.$$

  $\gamma_{i_{max}} \doteq 1$ for row with number $i_{max}$.

  The relaxation paarmeters for other rows are obtained by the formula:

  $$\gamma_i = \frac{1 - k_{i_{max}}}{1 - \varepsilon_i k_{i_{max}}} \qquad \left(\varepsilon_i = \frac{k_i}{k_{i_{max}}}\right).$$

- Balancing the right hand side:

  We use the notation:

  $$\frac{b_{i_{min}}}{a_{i_{min}i_{min}}} \doteq \min_{1\leq i\leq m} \frac{b_i}{a_{ii}}.$$

  $\gamma^*_{i_{min}} \doteq 1$ for row with number $i_{min}$.

  The relaxation parameters for other rows is obtained by the formula:

  $$\gamma^*_i = \frac{1}{\varepsilon^*_i} \qquad \left(\frac{b_i}{a_{ii}} = \varepsilon^*_i \frac{b_{i_{min}}}{a_{i_{min}i_{min}}}, \qquad i = 1,\ldots,m\right).$$

Table 1: Algorithm without balancing ($\gamma = 1$)

| $N$ | $\varepsilon$ | average $k$ | $t$ in *sec.* | D$X$ | $r_N$ | $\lvert(x,h) - \overline{X}_N\rvert$ |
|---|---|---|---|---|---|---|
| 10 000 | 3$e$-06 | 11 | 7.09 | 12.871 | 0.024 | 0.019 |
| 50 000 | 3$e$-06 | 11 | 34.84 | 12.832 | 0.011 | 0.016 |
| 115 000 | 3$e$-06 | 11 | 79.83 | 12.602 | 0.007 | 0.007 |

Table 2: Algorithm with balancing the iterative matrix

| $N$ | $\varepsilon$ | $k$ | $t$ in *sec.* | D$X$ | $r_N$ | $\lvert(x,h) - \overline{X}_N\rvert$ |
|---|---|---|---|---|---|---|
| 600 | 3$e$-01 | 67332 | 2153.09 | 11.084 | 0.092 | 0.074 |
| 700 | 3$e$-01 | 67332 | 2497.92 | 10.668 | 0.083 | 0.023 |
| 800 | 3$e$-01 | 67332 | 2848.84 | 10.971 | 0.079 | 0.020 |

But in practice the simultnaneous balancing is very difficult, because balancing one of the components puts very strong restrictions about choice for the elments of another component.

## 4.4 Balancing $L$ or $f$

Balancing only $L$ or $f$ is done with respect to the parameters of the given problem. The algorithm with balancing is compared with the algorithm without balancing for $\gamma = 1$. This valuse of the relaxation parameter is chosen, because then the rate if convergence is the highest. The algorithm with ablancing has lower efficiency, because it gives an inessential reduction of the variance as the length of the Markov chains increases considerably. This disadvantage can be overcome if the value of $\varepsilon$ is comparatively large. But then the systematic error increases. These are the reasons why the algorithm with balancing is not of great practical importance.

## 4.5 Numerical experiments

The algorithms are realized on FORTRAN 77 and the computational experiments are made with a real nonsymmetric general sparse square matrix of order 1107 which is diagonally dominant. The elements of the right hand side are random numbers in the interval $(0; 1)$, obtained with the generator URAND. The numerical experiments are obtained on work station HEWLETT PACKARD.

# 5 Chapter four. Monte Carlo algorithm for general sparse matrices

The main idea is that the elements of the matrix are stored in one dimensional array. The access to the elements is provided through the following arrays:

**NI** one dimensional array of order $m$, where $NI(i)$ is the number of nonzero elements in $i$-th row, $i = 1, \ldots, m$;

**A** one dimensional array of order $n = \sum_{i=1}^{m} NI(i)$, which contains all nonzero elements by rows;

**J** one dimensional array of order $n$, where $J(i)$ is the column index of $a_i \in A$.

When the presentation is used some difficulties appear. For example, the diagonal elements of the given system are necessary, but their indices in $A$ are unknown. They can be obtained by the formula:

$$diag_i = \{k : J(k) = i, k = IPS(i) + 1, \ldots, IPS(i) + i\}, \qquad i = 1, \ldots, m.$$

where $IPS$ is one dimensional array of order $m$ which contains partial sums of $NI(i)$:

$$\begin{aligned} IPS(1) &= 0 \\ IPS(i) &= IPS(i-1) + NI(i-1) \qquad i = 2, \ldots, m. \end{aligned}$$

# References


[1] J. H. Curtiss, *"Monte Carlo methods for the iteration of linear operators"*, J. Math Phys., vol. 32, pp. 209-232, **1954**.

[2] B. Dimitrov, *"Markov chains"*, Nauka and Art, Sofia, **1974**.

[3] I. T. Dimov, *"Minimization of the probable error for some Monte Carlo methods"*, Proc. International Conference on Mathematical Modelling and Scientific Computation, Varna, **1991**.

[4] I. T. Dimov, T. T. Dimov, T. V. Gurov, *"A new iterative Monte Carlo approach for inverse matrix problem"*, Journal of Computational and Applied Mathematics, vol. 92, No 4, pp. 15-35, **1998**.

[5] T. T. Dimov, *"Efficient Monte Carlo Algorithms for Inverting Matrices Arising in Mixed Finite Element Approximation"*, Proceedings of the Fourth International Conference on Numerical Methods and Applications, Sofia, August 1998, World Scientific, Singapore, New Jersey, London, Hong Kong, pp. 248-256.

[6] S M. Ermakov, G. A. Mikhailov, *"Statistical modelling"*, Nauka, Moskow, **1982**.

[7] G. H. Golub, Ch. F. Van Loan, *"Matrix Computations"*, The Johns Hopkins University Press, Baltimore and London.

[8] G. Megson, V. Aleksandrov, I. Dimov, *"Systolic matrix inversion using a Monte Carlo method"*, Journal of Parallel Algorithms and Applications, vol. 3, No 3/4, pp. 311-330, **1994**.

[9] N. Metropolis, S. M. Ulam, *"The Monte Carlo method"*, Journal of American Statistical Association, vol. 44, No 247, pp. 335-341, **1949**.

[10] I. M. Soboĺ, *"Monte Carlo numerical methods"*, Nauka, Moskow, **1973**.

[11] R. W. Wolff, *"Stochastic modelling and the theory of queues"*, Prentice-Hall International.