

***Fourier Analysis of Modified Nested Factorization  
Preconditioner for Three-Dimensional Isotropic  
Problems***

Pawan Kumar — Laura Grigori — Qiang Niu — Frederic Nataf

N° ?????

Janvier 2010

Thème NUM

 ***Rapport  
de recherche***



## Fourier Analysis of Modified Nested Factorization Preconditioner for Three-Dimensional Isotropic Problems

Pawan Kumar<sup>\*</sup>, Laura Grigori<sup>†</sup>, Qiang Niu<sup>‡</sup>, Frederic Nataf<sup>§</sup>

Thème NUM — Systèmes numériques  
Équipe-Projet Grand-large

Rapport de recherche n° ???? — Janvier 2010 — 25 pages

**Abstract:** For solving large sparse symmetric linear systems, arising from the discretization of elliptic problems, the preferred choice is the preconditioned conjugate gradient method. The convergence rate of this method mainly depends on the condition number of the preconditioner chosen. Using Fourier analysis the condition number estimate of common preconditioning techniques for two dimensional elliptic problem has been studied by Chan and Elman [SIAM Rev., 31 (1989), pp. 20-49]. Nested Factorization(NF) is one of the powerful preconditioners for systems arising from discretization of elliptic or hyperbolic partial differential equations. The observed convergence behavior of NF is better compared to well known ILU(0) or modified ILU. In this paper we introduce Modified Nested Factorization(MNF) which is an improvement over NF. It is proved that condition number of modified NF is  $O(h^{-1})$ . An optimal value of the parameter for the model problem is derived. The condition number of modified NF predicts the condition number of NF in limiting sense when the parameter is close to zero. Moreover it is proved that condition number of NF is atleast  $O(h^{-1})$ . Numerical results justify Fourier analytic method by exhibiting remarkable similarity in spectrum of periodic and Dirichlet problems.

**Key-words:** Nested Factorization, eigenvalues, eigenvectors, sparse LU, modified sparse LU, circulant matrices

<sup>\*</sup> INRIA Saclay - Ile de France, Laboratoire de Recherche en Informatique Université Paris-Sud 11, France (Email:pawan.kumar@lri.fr)

<sup>†</sup> INRIA Saclay - Ile de France, Laboratoire de Recherche en Informatique Université Paris-Sud 11, France (Email:laura.grigori@inria.fr)

<sup>‡</sup> School of Mathematical Sciences, Xiamen University, Xiamen, 361005, P.R. China; The work of this author was performed during his visit to INRIA, funded by China Scholarship Council; Email:kangniu@gmail.com

<sup>§</sup> Laboratoire J. L. Lions, CNRS UMR7598, Université Paris 6, France; Email:nataf@ann.jussieu.fr

## Analyse de Fourier des preconditioneurs du type factorisation emboite modifie pour les problemes isotropique en 3D

**Résumé :** Pour résoudre des grands systèmes linéaires d'équation symétrique obtenus de la discrétisation d'une équation aux dérivées partielles elliptique, on choisit le plus souvent la méthode du gradient conjugué préconditionné. La convergence de cette méthode dépend le plus souvent du conditionnement du système ainsi préconditionné. L'analyse de Fourier est une technique utilisée par Chan et Elman pour estimer le conditionnement de ce système préconditionné pour les problèmes 2D.

La factorisation emboîtée est un preconditionneur puissant pour les systèmes d'équation obtenus de la discrétisation d'une EDP elliptique ou hyperbolique. Les observations de la convergence de la factorisation emboîtée montrent qu'il se comporte mieux que ILU(0) ou ILU modifié. Dans ce papier on introduit la factorisation emboîtée qui est une amélioration de la factorisation emboîtée. Il est prouvé que le conditionnement de la factorisation emboîtée est de  $O(h^{-1})$ .

Les valeurs optimales des paramètres sont obtenues. Le conditionnement de la factorisation emboîtée modifiée permet de prédire celui de la factorisation emboîtée lorsque les paramètres sont près de zéro. De plus nous prouvons que le conditionnement du NF est au moins de  $O(h^{-1})$ . Les résultats numériques justifient les résultats de l'analyse de Fourier en exhibant avec des conditions remarquables sur des problèmes avec des conditions aux limites de Dirichlet et les conditions aux limites périodiques.

**Mots-clés :** factorization emboîtée, valeurs propres, vecteurs propres, LU creux, LU creux modifié, matrice circulaire

## 1 INTRODUCTION

For solving large sparse symmetric linear systems arising from the discretization of elliptic or hyperbolic partial differential equations, a preferred choice is the preconditioned conjugate gradient method. The convergence rate of this method mainly depends on the condition number of the preconditioned matrix; hence, the choice of a preconditioner is crucial. On the other hand, the choice of a preconditioner also depends on the cost of construction, storage requirements, and solve time. Although, most of these properties can be predicted in advance, predicting the condition number of a preconditioner is generally one of the most difficult properties to be determined in advance.

With fast setup time and very modest storage requirement, the Nested Factorization (NF) preconditioner introduced in [1] is one of the powerful preconditioners; it performs better compared to the widely used ILU(0) and Modified ILU [2] for certain class of problems [1, 3]. The method of NF differs from ILU(0) or Modified ILU(0) in that the preconditioning matrix in NF is not formed strictly from upper and lower triangular factors. Instead, block lower and upper factors are constructed using a procedure which adds one dimension at a time to the preconditioning matrix having the diagonal matrix on the lowest level. In modified NF (namely, MNF( $c$ )), a slight perturbation  $ch^2$  ( $c$  is constant and  $h$  is the mesh size independent of  $c$ ) is added to the diagonal, this is similar to the perturbation added to the diagonal of MILU preconditioner as suggested by Gustafsson [4].

The NF preconditioner has some important properties. If  $B_{NF}$  is the NF preconditioner, then  $\text{colsum}(B_{NF} - A) = 0$  (also known as zero colsum property), as a consequence the sum of the residuals in successive Krylov iterations remain zero, provided a suitable initial solution is used [1]. This property can provide a very useful check on the correctness of the implementation. Further, quoting [1], “the factorization procedure conserves material exactly for each phase at each linear iteration, and accommodates non-neighbor connections (arising from the treatment of the faults, completing the circle in three-dimensional coning studies, numerical aquifers, dual porosity/permeability systems etc.) in a natural way.” Moreover, for fluid flow problems it is proved in [5] that the lower and the upper triangular factors of NF are nonsingular. Due to these desirable qualities, the NF preconditioner is of particular interest in the oil reservoir industry; a method similar to NF is implemented in Schlumberger’s widely used Eclipse oil reservoir simulator [6].

Our goal in this paper is to give a condition number estimate of MNF preconditioned matrix using Fourier analysis. In principle, Fourier analysis is an exact analysis for the periodic problem. However, empirical observations suggest that the extreme eigenvalues of the periodic case remain in close agreement with the corresponding Dirichlet problem. So, the results obtained via Fourier analysis for the periodic case should predict the properties of the preconditioner for the corresponding Dirichlet case as well. It is in this light that most of the properties of the common preconditioners including Jacobi, Gauss-Siedel, SSOR, ADI, ILU(0), and MILU which were earlier obtained by “hard” analysis, were obtained easily and elegantly via Fourier analysis by Chan and Elman in [7]. For a two dimensional model problem, the method of NF is similar to block MILU, and Fourier analysis has been used to analyze the condition number for hyperbolic model problem [8]. Later, a similar technique was used to analyze

the condition number of ILU(0) and MILU for a three dimensional anisotropic model problem [9]. To the best of our knowledge, no results on the condition number of NF for three-dimensional case is known. Using Fourier analysis, we prove that the all the eigenvalues of NF preconditioned matrix are larger than one, and the condition number of MNF preconditioned matrix is of order  $h^{-1}$  which is similar to that of MILU. Although, our analysis is for MNF, in the limiting sense, i.e., when  $c$  tends to zero it essentially predicts the condition number of NF. Moreover, we will prove that the order of condition number for NF preconditioned matrix is at least  $O(h^{-1})$ . Finally, for an isotropic model problem we will propose an optimal value of the parameter.

Numerical results will illustrate the dependence of convergence of MNF on parameter  $c$ . The spectrum plots for different values of the parameter for the periodic problem are found to be in remarkable agreement with the corresponding Dirichlet problem, this justifies the applicability of our theoretical results for the corresponding Dirichlet problem.

The rest of this paper is organized as follows. In section 2, we briefly introduce some notations and collect some results on circulant matrices. In section 3, we describe the model problem and describe the MNF preconditioner. The Fourier eigenvalues for MNF is derived in this section. Later in section 4, we present various numerical experiments and compare the results of the periodic problem with that of Dirichlet problem for MNF and compare the obtained results with those of ILU and MILU. Section 5 concludes the paper. Finally, in appendix we present the proof of results leading to condition number estimate and the derivation of optimal value of the parameter.

## 2 MODEL PROBLEM AND THE PRECONDITIONER

We choose the same model as in [9] so that the results can be compared with ILU(0) and MILU. The model is the following three-dimensional anisotropic equation:

$$-(l_1 u_{xx} + l_2 u_{yy} + l_3 u_{zz}) = r \quad (1)$$

defined on a unit cube  $\Omega = \{0 \leq x, y, z \leq 1\}$ , with  $l_1, l_2, l_3 \geq 0$ , and with the periodic boundary conditions as follows

$$\begin{aligned} u(x, y, 0) &= u(x, y, 1), \\ u(x, 0, z) &= u(x, 1, z), \\ u(0, y, z) &= u(1, y, z). \end{aligned}$$

The discretization scheme considered in the interior of the domain is the second order finite differences on a uniform  $n \times n \times n$  grid, with mesh size  $h = \frac{1}{n+1}$  along  $x$ ,  $y$ , and  $z$  directions. Here we shall use the notation  $h$  to denote the mesh size for the periodic case. With this discretization we get a system of equation

$$Au = b. \quad (2)$$

It is useful to express the matrix  $A$  arising from the periodic boundary conditions using the notation of circulant matrices and the Kronecker product. We introduce these notations as follows.

**Definition 2.1** Let  $C$  be a matrix of size  $pq \times pq$ . We call  $C$  a **block circulant matrix** if it has the following form

$$C = \text{Bcirc}_p(C_0, C_{p-1}, \dots, C_2, C_1) = \begin{pmatrix} C_0 & C_{p-1} & \cdots & C_2 & C_1 \\ C_1 & C_0 & C_{p-1} & \ddots & \vdots \\ \vdots & C_1 & C_0 & \ddots & \vdots \\ C_{p-2} & & \ddots & \ddots & C_{p-1} \\ C_{p-1} & C_{p-2} & \cdots & C_1 & C_0 \end{pmatrix}_{pq \times pq},$$

where each of the blocks  $C_i$  are matrices of size  $q \times q$  each. We observe that a block circulant matrix is completely specified by a block row. However if  $q = 1$ , then we simply call it **circulant matrix** and denote it by  $\text{circ}_p(C_0, C_{p-1}, \dots, C_2, C_1)$ .

**Notation 2.2** Further, for **block circulant tridiagonal matrices** we introduce the following notation

$$\text{Bctrid}_p(C_2, C_0, C_1) = \begin{pmatrix} C_0 & C_1 & & & C_2 \\ C_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & C_0 & C_1 \\ C_1 & & & C_2 & C_0 \end{pmatrix}_{pq \times pq},$$

where each of the blocks  $C_i$  are matrices of size  $q \times q$  each. However if  $q = 1$ , then we denote it by  $\text{ctrid}_p(C_2, C_0, C_1)$ .

**Notation 2.3** For **block tridiagonal matrix** with constant block bands we introduce the following notation

$$\text{Btrid}_p(F_2, F_0, F_1) = \begin{pmatrix} F_0 & F_1 & & & \\ F_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & F_0 & F_1 \\ & & & F_2 & F_0 \end{pmatrix}_{pq \times pq},$$

where each of the blocks  $F_i$  are matrices of size  $q \times q$  each. If  $q = 1$ , then we simply denote it by  $\text{trid}_p(F_2, F_0, F_1)$ .

**Definition 2.4** The Kronecker product  $\otimes$  is an operation on two matrices of arbitrary size resulting in a block matrix. Let  $A = (a_{i,j})$  and  $B = (b_{i,j})$ , then by  $A \otimes B$  we mean

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ \vdots & \ddots & & \vdots \\ \vdots & \cdots & & \vdots \\ a_{n1}B & a_{n2}B & \cdots & a_{nn}B \end{pmatrix}.$$

If the difference operators are scaled by step size  $h^2$ , then equation of (2) corresponding to the  $(i, j, k)^{th}$  grid point is the following:

$$\begin{aligned} a_{i,j,k}u_{i,j,k} + b_{i,j,k}u_{i+1,j,k} + c_{i,j,k}u_{i,j+1,k} + d_{i,j,k}u_{i-1,j,k} \\ + e_{i,j,k}u_{i,j-1,k} + f_{i,j,k}u_{i,j,k+1} + g_{i,j,k}u_{i,j,k-1} = w_{i,j,k}, \end{aligned} \quad (3)$$

where  $1 \leq i, j, k \leq n$ , and

$$\begin{aligned} b_{i,j,k} &= 0, \quad i = n, \\ c_{i,j,k} &= 0, \quad j = n, \\ f_{i,j,k} &= 0, \quad k = n, \\ d_{i,j,k} &= 0, \quad i = 1, \\ e_{i,j,k} &= 0, \quad j = 1, \\ g_{i,j,k} &= 0, \quad k = 1. \end{aligned} \quad (4)$$

For an anisotropic model problem, we have the following assignments:

$$\begin{aligned} a_{i,j,k} &= 2(l_1 + l_2 + l_3), \\ b_{i,j,k} &= -l_1, \\ c_{i,j,k} &= -l_2, \\ d_{i,j,k} &= -l_1, \\ e_{i,j,k} &= -l_2, \\ f_{i,j,k} &= -l_3, \\ g_{i,j,k} &= -l_3, \end{aligned} \quad (5)$$

where  $w_{i,j,k} = h^2 r(i, j, k)$ . Here the subscript  $(i, j, k)$  correspond to the grid location  $(ih, jh, kh)$ .

Let  $I_k$  denote the identity matrix of size  $k \times k$ . Using the notation of circulant matrix and the Kronecker product, the coefficient matrix corresponding to formula (3) is expressed as follows

$$\begin{aligned} A &= Bctrid_n \left( -l_3 I_{n^2}, \widehat{D}, -l_3 I_{n^2} \right), \\ \widehat{D} &= Bctrid_n \left( -l_2 I_n, \overline{D}, -l_2 I_n \right), \\ \overline{D} &= ctrid_n \left( -l_1, d, -l_1 \right). \end{aligned}$$

We consider now the same problem (1) with the following Dirichlet boundary condition

$$\begin{aligned} u(x, y, 0) &= 0, \\ u(x, 0, z) &= 0, \\ u(0, y, z) &= 0. \end{aligned} \quad (6)$$

To differentiate the Dirichlet problem with that of periodic problem, we shall use bold face letters to denote the matrices corresponding to the Dirichlet case. Using second order finite differences with the Dirichlet boundary conditions 7 above, we obtain the matrix  $\mathbf{A}$  corresponding to the Dirichlet case as follows

$$\mathbf{A} = \mathbf{D} + \mathbf{L}_1 + \mathbf{L}_1^T + \mathbf{L}_2 + \mathbf{L}_2^T + \mathbf{L}_3 + \mathbf{L}_3^T,$$

where

$$\begin{aligned}\mathbf{L}_3 &= Btrid_n(-l_3 I_{n^2}, 0, 0), \\ \mathbf{L}_2 &= I_n \otimes Btrid_n(-l_2 I_n, 0, 0), \\ \mathbf{L}_1 &= I_{n^2} \otimes trid_n(-l_1, 0, 0).\end{aligned}$$

For the above model problem the nested factorization preconditioner  $\mathbf{B}$  for the Dirichlet problem is defined as follows:

$$\begin{aligned}\mathbf{B} &= (\mathbf{P} + \mathbf{L}_3) (\mathbf{I} + \mathbf{P}^{-1} \mathbf{L}_3^T), \\ \mathbf{P} &= (\mathbf{T} + \mathbf{L}_2) (\mathbf{I} + \mathbf{T}^{-1} \mathbf{L}_2^T), \\ \mathbf{T} &= (\mathbf{M} + \mathbf{L}_1) (\mathbf{I} + \mathbf{M}^{-1} \mathbf{L}_1^T),\end{aligned}\tag{7}$$

where  $\mathbf{M} = \text{diag}(\mathbf{A}) - \mathbf{L}_1 \mathbf{M}^{-1} \mathbf{L}_1^T - \text{colsum}(\mathbf{L}_2 \mathbf{T}^{-1} \mathbf{L}_2^T) - \text{colsum}(\mathbf{L}_3 \mathbf{P}^{-1} \mathbf{L}_3^T)$ .

Here we denote  $\text{colsum}(K)$  to mean diagonal matrix  $K$  formed from the vector  $\mathbf{1}K$ , here  $\mathbf{1}$  stands for vector of all ones; and by  $\text{diag}(K)$  we mean the strict diagonal of matrix  $K$ .

The MNF preconditioner has the same hierarchical definition as 7 above; it differs from NF in that the diagonal matrix  $M$  above is replaced by

$$\mathbf{M}_{\text{MNF}} = \text{diag}(\mathbf{A}) + ch^2 - \mathbf{L}_1 \mathbf{M}^{-1} \mathbf{L}_1^T - \text{colsum}(\mathbf{L}_2 \mathbf{T}^{-1} \mathbf{L}_2^T) - \text{colsum}(\mathbf{L}_3 \mathbf{P}^{-1} \mathbf{L}_3^T),\tag{8}$$

where  $ch^2$  is a perturbation added to the diagonal matrix  $\mathbf{M}$ . Here the constant  $c$  is independent of the mesh size  $h$ . We will use notations  $\tilde{B}$ ,  $\tilde{P}$ , and  $\tilde{T}$  for MNF preconditioner corresponding to the periodic case.

The construction and the solution procedure for MNF is very similar to NF, see [1, 3].

### 3 FOURIER ANALYSIS OF THE MNF PRE-CONDITIONER

In this section, we will derive the Fourier eigenvalues of the MNF preconditioned matrix. For clarity and simplicity we restrict our analysis to the isotropic problem ( $l_1 = l_2 = l_3 = 1$ ), however, similar analysis holds for the general anisotropic case. We shall exhibit numerical comparisons and results for the general anisotropic case. Along the way we will outline certain assumptions on which our analysis will be based. These assumptions are similar to those made in [7] and has been justified their appropriately.

We shall treat matrices  $\tilde{B}$  and  $A$  as if they were periodic.

Fourier analysis is an exact analysis only for constant coefficient matrix with periodic boundary conditions. Our original matrix  $A$  is indeed a constant coefficient matrix but the corresponding MNF preconditioner is not a constant coefficient matrix since the diagonal matrix  $\tilde{M}$  is not constant; the recursive expression for  $\tilde{M}$  leads to varying entries in  $\tilde{M}$ . However, the values of  $\tilde{M}$  away from the boundary tend to certain value in limiting sense (apparent for a large matrix). In other words, for a large size matrix most of the entries of  $\tilde{M}$  are close to entries of  $\tilde{M}$  near the middle entry, i.e., the entry  $\tilde{M}_{n^3/2, n^3/2}$ . To observe this we plot a histogram of the entries of  $\tilde{M}$  corresponding to the Dirichlet boundary condition for  $120 \times 120 \times 120$  grid in Figure (1). We find that most of the entries

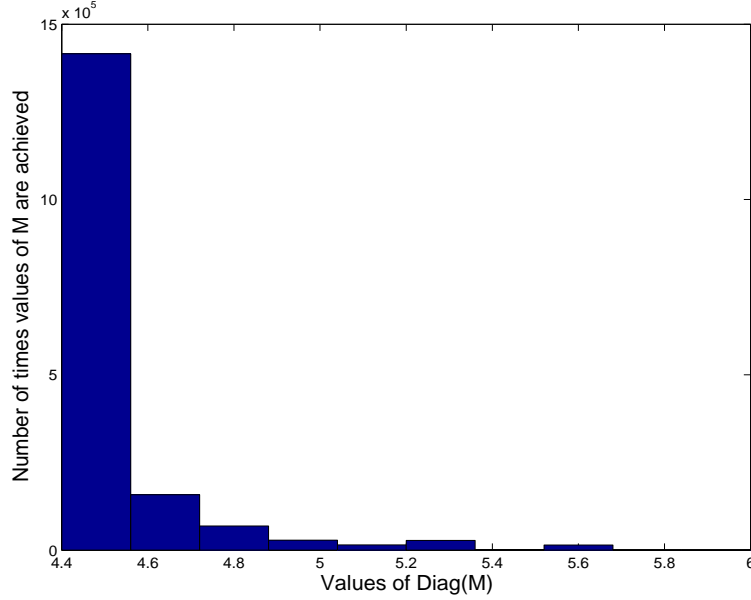


Figure 1: Histogram plot of  $\text{Diag}(\tilde{M})$  for  $\text{MNF}(c_d)$ , Here  $c_d = 1.45\pi^2$  is the optimal value of the parameter  $c$  for the Dirichlet case as derived in appendix. The matrix corresponds to 7-point discretization scheme applied to (1) on  $120 \times 120 \times 120$  unit cube with coefficients set such that the matrix is symmetric with lower bands  $(l_1, l_2, l_3) = (1, 1, 1)$

of  $\tilde{M}$  are close to the middle entry of  $\tilde{M}$ . So in some sense we can treat the matrix  $\tilde{M}$  to be a constant coefficient matrix for large dimension. To find this constant value for the periodic case, we force the diagonal matrix  $M_{MNF}$  to be constant in recursion (8) in the previous section. But this recursion will lead to a fifth degree equation which may not be solved using radicals. Instead, we observe that the matrix  $\tilde{T}$  for MNF preconditioner is

$$\begin{aligned} \tilde{T} &= M_{MNF} + L_1 + L_1^T + L_1 M_{MNF}^{-1} L_1^T, \\ &= \tilde{M} + L_1 + L_1^T, \end{aligned}$$

where

$$\tilde{M} = \text{diag}(A) + ch^2 - \text{colsum}(L_2 \tilde{T}^{-1} L_2^T) - \text{colsum}(L_3 \tilde{P}^{-1} L_3^T). \quad (9)$$

Using the notation of circulant matrix and the Kronecker product, the MNF preconditioner is now defined as follows:

$$\begin{aligned}
\tilde{B} &= (\tilde{P} + L_3)(I + \tilde{P}^{-1}L_3^T), \quad \tilde{P} \text{ is of size } n^3 \times n^3, \\
L_3 &= \text{Bcirc}_n(0, \dots, 0, -l_3 I_{n^2}), \\
L_3^T &= \text{Bcirc}_n(0, -l_3 I_{n^2}, 0, \dots, 0), \\
\tilde{P} &= I_n \otimes P_0, \\
P_0 &= (\hat{T} + \hat{L}_2)(I + \hat{T}^{-1}\hat{L}_2^T), \\
\hat{L}_2 &= \text{Bcirc}_n(0, \dots, 0, -l_2 I_n), \\
\hat{L}_2^T &= \text{Bcirc}_n(0, -l_2 I_n, 0, \dots, 0), \\
\hat{T} &= I_n \otimes T_0, \quad T_0 \text{ is of size } n \times n, \\
T_0 &= \text{circ}_n(\tilde{m}, -l_1, 0, \dots, 0, -l_1), \\
\tilde{m} &= d + ch^2 - \frac{l_2^2}{\tilde{m} - 2l_1} - \frac{l_3^2}{\tilde{m} - 2l_1 - 2l_2 + \frac{l_2^2}{\tilde{m} - 2l_1}}. \quad (10)
\end{aligned}$$

We notice here that recursion (10) is obtained from recursion (9); the recursion (10) is satisfied by the roots of a fourth degree equation which can now be solved by radicals.

To obtain recursion (10), we observe that the matrices  $L_2$  and  $L_2^T$  being constant coefficient circulant matrices and  $\tilde{T}$  being a circulant matrix, we have

$$\text{colsum}(L_2 \tilde{T}^{-1} L_2^T) = \frac{l_2^2}{\text{colsum}(\tilde{T})} = \frac{l_2^2}{\tilde{m} - 2l_1}.$$

Also,  $L_3$  and  $L_3^T$  being constant coefficient circulant matrices and  $\tilde{P}$  being a circulant matrix, we have

$$\text{colsum}(L_3 \tilde{P}^{-1} L_3^T) = \frac{l_3^2}{\text{colsum}(\tilde{P})} = \frac{l_3^2}{\text{colsum}(\hat{T} + L_2 + L_2^T) + \text{colsum}(L_2 \tilde{T}^{-1} L_2^T)}.$$

It is easy to see that  $\text{colsum}(\hat{T} + L_2 + L_2^T) = \tilde{m} - 2l_1 - 2l_2$ .

The recurrence (10) is satisfied by the roots of the following fourth degree equation

$$x^4 + c_1 x^3 + c_2 x^2 + c_3 x + c_4 = 0,$$

where  $c_1 = -14 - ch^2$ ,  $c_2 = 71 + 8ch^2$ ,  $c_3 = -154 - 21ch^2$ , and  $c_4 = 121 + 18ch^2$ . In this fourth degree equation, we choose a root with maximum magnitude

$$\begin{aligned}
\tilde{m} &= 7/2 + 1/4 ch^2 + 1/4 \sqrt{f_h} + \\
&1/4 \sqrt{2} \sqrt{\frac{10 \sqrt{f_h} + 8 ch^2 \sqrt{f_h} + c^2 h^4 \sqrt{f_h} + 24 ch^2 + 10 c^2 h^4 + c^3 h^6}{\sqrt{f_h}}}, \quad (11)
\end{aligned}$$

where  $f_h = 4ch^2 + c^2 h^4$ . For the isotropic case, the recurrence relation can have four roots, the maximum of these roots in magnitude approximates the value most achieved, as suggested by the histogram plot of diagonal matrix  $\tilde{M}$  for the Dirichlet case in Figure 1. In this figure,  $c_p$  is the optimal value of parameter for the periodic case as derived in the appendix.

According to an argument in [7], the extreme eigenvalues for the periodic and the corresponding Dirichlet problems are same provided  $n = 2n_d$ . Here  $n_d + 1 = 1/h_d$  and  $h_d$  is the mesh size for the Dirichlet problem. This assumption is important for our analysis as we are interested in studying the condition number of the preconditioned matrix, and the extreme eigenvalues play an important role.

With the above assumptions, we have obtained a periodic constant coefficient MNF preconditioner, for which exact Fourier analysis can be used.

Eigenvectors of  $A$  are found by applying the operator  $A$  to eigenvectors  $v^{s,t,r}$ . The  $(i, j, k)^{th}$  grid component of eigenvector  $v^{s,t,r}$  is given by

$$v_{i,j,k}^{s,t,r} = e^{i\theta_s} e^{j\phi_t} e^{k\xi_r}, \quad (12)$$

where  $\iota = \sqrt{-1}$ ,  $\theta_s = \frac{2\pi}{n+1}s$ ,  $\phi_t = \frac{2\pi}{n+1}t$ , and  $\xi_r = \frac{2\pi}{n+1}r$ , for  $r, s, t = 1, \dots, n$ . The eigenvalue  $\lambda_{s,t,r}(A)$  of the matrix  $A$  is determined by substituting (12) for  $u_{i,j,k}$  in the left hand side of (3) and it is found to be

$$\lambda_{s,t,r}(A) = 4 \left( l_1 \sin^2 \frac{\theta_s}{2} + l_2 \sin^2 \frac{\phi_t}{2} + l_3 \sin^2 \frac{\xi_r}{2} \right). \quad (13)$$

For circulant matrices following results hold.

**Lemma 3.1** [10] *Any circulant matrix of size  $n$  share the same set of eigenvectors.*

Using lemma (3.1) above, we have the following result.

**Lemma 3.2** *Let  $S$  and  $R$  be two given circulant matrices with eigenvalues  $\lambda_{s,t,r}(S)$  and  $\lambda_{s,t,r}(R)$  respectively. Then the eigenvalues of  $S + R$  and  $SR$  corresponding to the  $(s, t, r)^{th}$  grid point is given as follows:*

- $\lambda_{s,t,r}(S + R) = \lambda_{s,t,r}(S) + \lambda_{s,t,r}(R)$ .
- $\lambda_{s,t,r}(SR) = \lambda_{s,t,r}(S)\lambda_{s,t,r}(R)$ .

**Proof:** It follows easily using lemma (3.1) above.

Using the lemma 3.2 above, eigenvalue  $\lambda_{s,t,r}(\tilde{B}^{-1}A)$  of MNF preconditioned matrix is then given by

$$\lambda_{s,t,r}(\tilde{B}^{-1}A) = \frac{\lambda_{s,t,r}(A)}{\lambda_{s,t,r}(\tilde{B})}, \quad (14)$$

where  $\lambda_{s,t,r}(\tilde{B})$  is given hierarchically as follows:

$$\begin{aligned}
\lambda_{s,t,r}(\tilde{B}) &= (\lambda_{s,t,r}(\tilde{P}) + \lambda_{s,t,r}(L_3)) \left( 1 + \frac{\lambda_{s,t,r}(L_3^T)}{\lambda_{s,t,r}(\tilde{P})} \right), \\
\lambda_{s,t,r}(\tilde{P}) &= (\lambda_{s,t,r}(\tilde{T}) + \lambda_{s,t,r}(L_2)) \left( 1 + \frac{\lambda_{s,t,r}(L_2^T)}{\lambda_{s,t,r}(\tilde{T})} \right), \\
\lambda_{s,t,r}(\tilde{T}) &= \lambda_{s,t,r}(\tilde{M}) + \lambda_{s,t,r}(L_1) + \lambda_{s,t,r}(L_1^T), \\
\lambda_{s,t,r}(L_1) &= -l_1 e^{\iota \theta_s}, \\
\lambda_{s,t,r}(L_1^T) &= -l_1 e^{-\iota \theta_s}, \\
\lambda_{s,t,r}(\tilde{M}) &= \tilde{m}, \\
\lambda_{s,t,r}(L_2) &= -l_2 e^{\iota \phi_t}, \\
\lambda_{s,t,r}(L_2^T) &= -l_2 e^{-\iota \phi_t}, \\
\lambda_{s,t,r}(L_3) &= -l_3 e^{\iota \xi_r}, \\
\lambda_{s,t,r}(L_3^T) &= -l_3 e^{-\iota \xi_r}.
\end{aligned} \tag{15}$$

The eigenvalues for  $L_1$ ,  $L_2$ ,  $L_3$ ,  $U_1$ ,  $U_2$ ,  $U_3$ , and  $\tilde{M}$  were found by inspection, for instance, if (3) denotes the stencil for the original matrix  $A$ , then the stencils (or equations) for the matrices  $L_1, L_2, L_3, L_1^T, L_2^T, L_3^T$ , and  $\tilde{M}$  are given by

$$\begin{aligned}
\text{stencil for } \tilde{M} &= \tilde{m} u_{i,j,k}, \\
\text{stencil for } L_1 &= -l_1 u_{i-1,j,k}, \\
\text{stencil for } L_1^T &= -l_1 u_{i+1,j,k}, \\
\text{stencil for } L_2 &= -l_2 u_{i,j-1,k}, \\
\text{stencil for } L_2^T &= -l_2 u_{i,j+1,k}, \\
\text{stencil for } L_3 &= -l_3 u_{i,j,k-1}, \\
\text{stencil for } L_3^T &= -l_3 u_{i,j,k+1}.
\end{aligned} \tag{16}$$

After substituting the eigenvector (12) in (16) for  $u_{i,j,k}$ , a straightforward computation gives the required eigenvalues in (15).

Consider now the isotropic problem ( $l_1 = l_2 = l_3 = 1$ ). From the expression for the eigenvalues of MNF preconditioned matrix, we obtain the following estimate for the condition number.

**Theorem 3.3** *If  $0 < c < \frac{16\pi^2}{45} (35 + 15\sqrt{5})$ , then for MNF preconditioned isotropic operator we have*

$$\kappa(\tilde{B}^{-1}A) = O(h^{-1}),$$

and for NF preconditioned isotropic operator, we have

$$\kappa(B^{-1}A) \geq O(h^{-1}). \tag{17}$$

## 4 NUMERICAL EXPERIMENTS

The aim of this section is two folds:

- To present the convergence results of MNF with the optimal value of the parameter  $c_{opt} = \frac{40}{9}\pi^2$ , and to compare the results with other common preconditioners including NF.
- To exhibit remarkable similarity in the spectrum distribution of MNF preconditioned matrix for periodic and the corresponding Dirichlet problems. For this, we will compare the extreme eigenvalues and the condition number for different values of parameter for the periodic and the corresponding Dirichlet preconditioned matrices.

All numerical experiments are performed in double precision arithmetic in MATLAB except the MNF solver subroutine which is implemented in FORTRAN 90. For solving the system  $Au = b$  for the Dirichlet problem via PCG, the initial solution vector is chosen to be vector of all zeros. The known solution vector is chosen to be a random vector. The stopping criteria for solving the iteration is the decrease of relative residual below  $10^{-12}$ , and the maximum number of iterations allowed is 200. The eigenvalues for MNF preconditioned Dirichlet problem for matrices of size  $n_d \geq 16$  is approximated by the harmonic Ritz values obtained after 20 steps of the Arnoldi iteration [11].

### 4.1 Performance of MNF

For numerical experiments we consider the following test cases:

1. Data Set 1 :  $(l_1, l_2, l_3) = (1, 1, 1)$
2. Data Set 2 :  $(l_1, l_2, l_3) = (1, 1, 0.01)$
3. Data Set 3 :  $(l_1, l_2, l_3) = (1, 0.01, 0.01)$

The test cases above are same as those found in [9]. In Table (1), convergence results are shown for Data set 1. The optimal value of the parameter chosen for MNF is  $c_d = 1.45\pi^2$ , where  $c_d$  is the optimal parameter for the Dirichlet case as derived in the appendix. We observe that MNF performs better compared to NF itself, and the difference in number of steps for convergence becomes significant as the problem size increases. On the other hand, MILU performs better than ILU but it is slower when compared to both NF and MNF.

### 4.2 Numerical Experiments with Fourier Eigenvalues

In Figures (5), (6), and (7) we present the spectrum comparison for Dirichlet and periodic case for  $n_d = 10$  for MNF preconditioned matrix. In Figure (8), we have a similar plot for NF. We observe that the similarity in the spectrum is remarkable, apart from coherence in extreme eigenvalues, the clustering traits found in Figures (6) and (7) for the Dirichlet problem is captured extremely well by the periodic problem. To verify such coherence in extreme eigenvalues for large problem sizes we do the following experiment.

In Tables (2), (3), and (4) we compare the extreme eigenvalues for large matrices, we tabulate the minimum, maximum eigenvalues, and the condition

Table 1: Convergence results for MNF( $c_d$ ) PCG, Data set 1

$h_d$	ILU(0)		MILU		NF		MNF( $c_p$ )	
	iter	err	iter	err	iter	err	iter	err
$\frac{1}{16}$	30	e-12	31	e-13	16	e-13	14	e-12
$\frac{1}{32}$	55	e-11	46	e-12	23	e-12	20	e-12
$\frac{1}{64}$	128	e-10	74	e-11	33	e-12	28	e-12
$\frac{1}{120}$	169	e-11	98	e-12	46	e-12	38	e-12

number of  $\tilde{B}^{-1}A$ . We observe that the extreme eigenvalues and the condition numbers for the periodic case are in close agreement with Dirichlet problem. For data set 1 (Table (2)), we find that the minimum eigenvalues for the periodic and Dirichlet problem tend to one as the matrix size increases; the minimum eigenvalue of one is indeed achieved for periodic problem corresponding to the grid size  $240 \times 240 \times 240$ . This is in agreement with theorem 5.4 which is true for small enough grid size  $h$ . Comparing with data in [9] for MILU, we find that the condition number of MNF preconditioned matrix is better, and comparatively the eigenvalues are closer to one.

In figures (2), (3), and (4) the maximum eigenvalue, the minimum eigenvalue, and the condition number of  $B_{MNF}^{-1}A$  for  $n_d = 64$  for different values of parameter  $c$  are shown. We observe that the matching of condition numbers for the periodic and Dirichlet case seems to be perfect for the value of  $c$  larger than its predicted optimal value.

The optimal value of the parameter for the periodic and the corresponding Dirichlet problem are

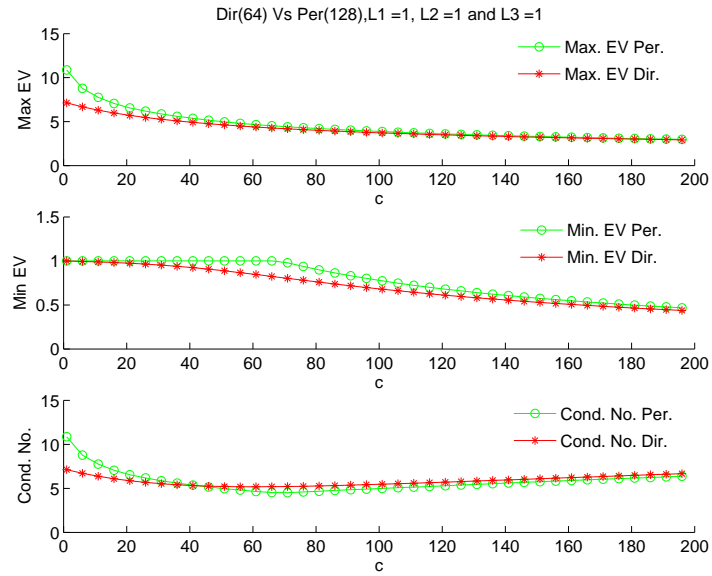
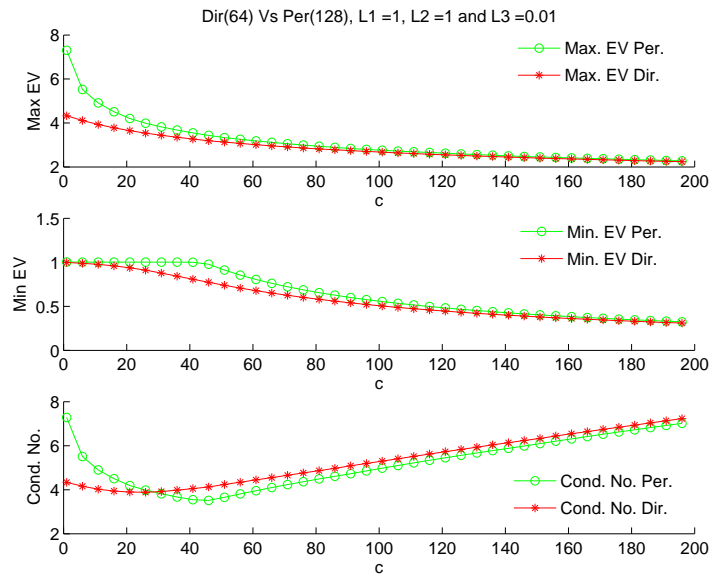
$$\begin{aligned} c_p &= 5.8\pi^2, \\ c_d &= 1.45\pi^2, \end{aligned}$$

respectively, as derived in appendix for Data Set 1. These optimal values are predicted in our plot for Data Set 1, see Figure (2) for example. The predicted optimal value is a slight underestimate of the exact optimal value observed in the plots, this is because of the approximation in the condition number while finding the optimal value (See appendix). Nevertheless, the predicted optimal value remains a good “heuristic“ estimate, given that the flatness in the graph of condition number begin to appear immediately after this optimal value.

On the other hand, in Figure (3) corresponding to Data Set 2, we observe that the condition number is more sensitive with parameter  $c$ , and the flatness around the optimal value is not as prominent as found in Data Set 1. From these plots it is also clear that the optimal value of parameter is problem dependent.

## 5 CONCLUSION

We have introduced modified nested factorization (MNF) which is an improved version of popular nested factorization preconditioner. We have implemented a tool of Fourier analysis to analyse MNF along with the nested factorization preconditioner. We derive that the condition number of MNF preconditioned matrix is  $O(h^{-1})$ . The analysis for MNF also predicts the condition number

Figure 2: Condition No. versus  $c$ ,  $n_d = 64$  for Data set 1Figure 3: Condition No. versus  $c$ ,  $n_d = 64$  for Data set 2

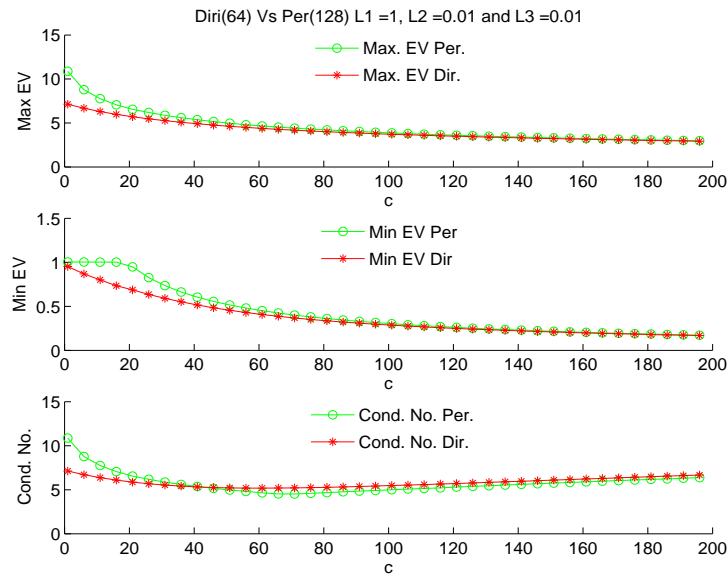


Figure 4: Condition No. versus  $c$ ,  $n_d = 64$  for Data set 3

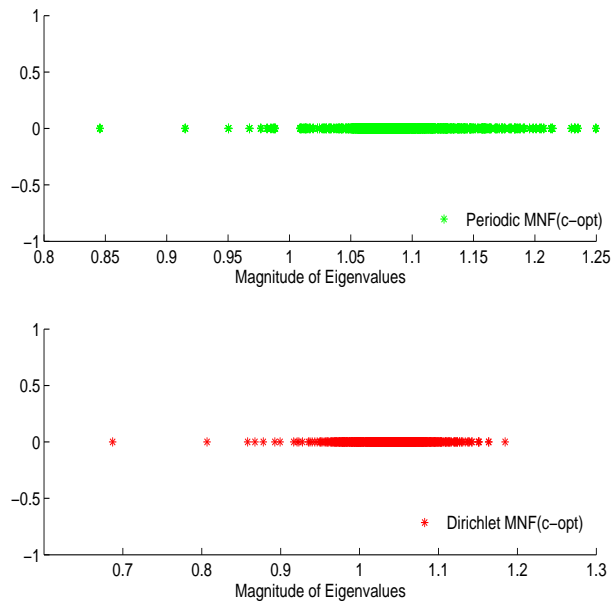


Figure 5: Spectrum comparison  $MNF(c_p)$  pre. matrix for Data set 1

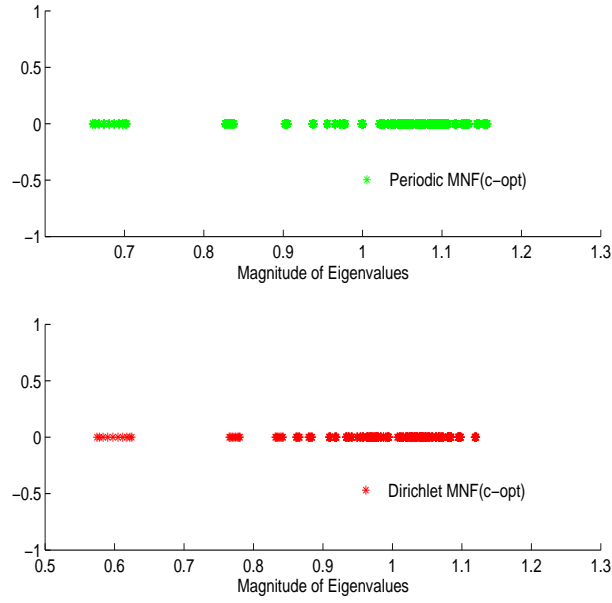
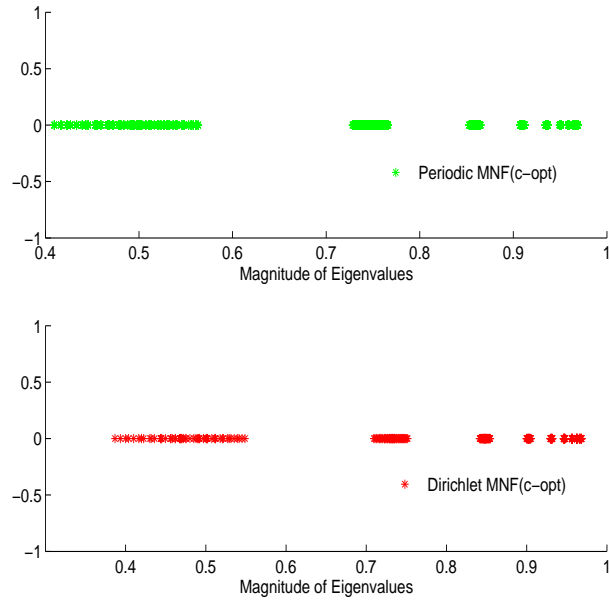
Figure 6: Spectrum comparison  $MNF(c_p)$  pre. matrix for Data set 2Figure 7: Spectrum comparison  $MNF(c_p)$  pre. matrix for Data set 3

Table 2: Dir. Versus Per. MNF( $c_p$ ) for Data set 1

$h_d$	$\lambda_{min}$		$\lambda_{max}$		cond. no.	
	Dir.	Per.	Dir.	Per.	Dir.	Per.
$\frac{1}{8}$	0.663	0.798	1.098	1.136	1.655	1.423
$\frac{1}{16}$	0.738	0.948	1.503	1.632	2.035	1.720
$\frac{1}{32}$	0.664	0.840	2.166	2.356	3.261	2.803
$\frac{1}{64}$	0.906	0.906	4.020	4.226	5.267	4.660
$\frac{1}{120}$	0.928	1.000	8.292	8.794	8.934	8.793

Table 3: Dir. Versus Per. MNF( $c_p$ ) for Data set 2

$h_d$	$\lambda_{min}$		$\lambda_{max}$		cond. no.	
	Dir.	Per.	Dir.	Per.	Dir.	Per.
$\frac{1}{8}$	0.555	0.635	1.057	1.071	1.903	1.687
$\frac{1}{16}$	0.613	0.708	1.354	1.436	2.208	2.026
$\frac{1}{32}$	0.667	0.781	2.002	2.123	2.997	2.715
$\frac{1}{64}$	0.734	0.894	3.106	3.310	4.229	3.701
$\frac{1}{120}$	0.787	1.001	4.936	5.283	6.264	5.277

Table 4: Dir. Versus Per. MNF( $c_p$ ) for Data set 3

$h_d$	$\lambda_{min}$		$\lambda_{max}$		cond. no.	
	Dir.	Per.	Dir.	Per.	Dir.	Per.
$\frac{1}{8}$	0.375	0.402	0.950	0.951	2.533	2.365
$\frac{1}{16}$	0.404	0.422	0.980	0.988	2.422	2.338
$\frac{1}{32}$	0.423	0.449	0.995	0.998	2.352	2.219
$\frac{1}{64}$	0.447	0.504	1.037	1.060	2.319	2.101
$\frac{1}{120}$	0.573	0.483	1.418	1.325	2.473	2.740

of nested factorization in the “limiting sense”. Moreover, from our analysis of an isotropic model problem, it is established that for nested factorization preconditioned matrix, the order of condition number is at least  $O(h^{-1})$ .

Future work may involve studying a dynamically modified version of NF where the parameter is chosen dynamically as in [13] for Modified ILU.

**Appendix.** In this section, we will give a condition number estimate for modified nested factorization as stated in Theorem 3.3. In the end, we will derive an optimal value for parameter  $c$ .

A symbolic algebra software MAPLE is used to perform some algebraic manipulation. We tabulate relevant built in functions. Let  $X = (x_1, x_2, \dots, x_n)$  where  $x_i, i = 1 \dots n$  are independent variables.

In Table (5) the maximize and minimize functions are the not the numerical maximum of minimum, but in fact they are the actual extreme values found (or proved) symbolically by MAPLE for the specified domain of  $X$ .

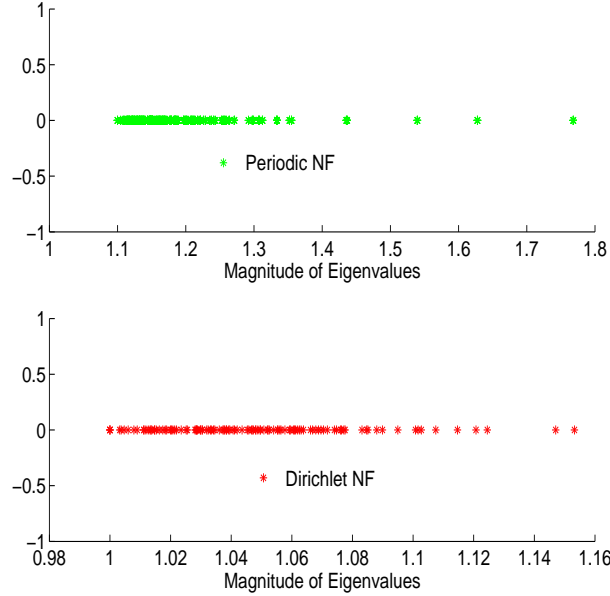


Figure 8: Spectrum comparison NF pre. matrix for Data set 1

Table 5: MAPLE functions

Syntax	Result
$\text{simplify}(f(X))$	Simplified expression
$\text{numer}\left(\frac{N(X)}{D(X)}\right)$	$N(X)$
$\text{denom}\left(\frac{N(X)}{D(X)}\right)$	$D(X)$
$\text{collect}(f(X), x_i)$	Collect like powers of $f$ in $x_i$
$\text{expand}(f(X))$	Expand the expression $f$
$\text{minimize}(f(X), x_1 = r_1^1 \dots r_2^1, \dots, x_n = r_1^n \dots r_2^n)$	Find the minimum of $f$ for given range.
$\text{maximize}(f(X), x_1 = r_1^1 \dots r_2^1, \dots, x_n = r_1^n \dots r_2^n)$	Find the maximum of $f$ for given range.

The Fourier eigenvalues for the coefficient matrix  $A$ , and MNF preconditioner  $\tilde{B}$  are given as follows:

$$\begin{aligned} \lambda_{s,t,r}(A) &= 4 \left( \sin^2 \frac{\theta_s}{2} + \sin^2 \frac{\phi_t}{2} + \sin^2 \frac{\xi_r}{2} \right), \\ \lambda_{s,t,r}(\tilde{B}) &= \lambda_{s,t}(\tilde{P}) + \frac{1}{\lambda_{s,t}(\tilde{P})} - 2\cos(\xi_r), \\ \lambda_{s,t}(\tilde{P}) &= \lambda_s(\tilde{T}) + \frac{1}{\lambda_s(\tilde{T})} - 2\cos(\phi_t), \\ \lambda_s(\tilde{T}) &= \tilde{m} - 2\cos(\theta_s), \\ \tilde{m} &\approx \frac{7 + \sqrt{5}}{2} + Kh, \end{aligned}$$

where  $K = \frac{3\sqrt{5}\sqrt{c}}{10}$ . Due to the periodicity of eigenvalues, i.e.,

$$\begin{aligned}\lambda_{s,t,r}(A) |_{(\theta_s, \phi_t, \xi_r)} &= \lambda_{s,t,r}(A) |_{(2\pi-\theta_s, 2\pi-\phi_t, 2\pi-\xi_r)}, \\ \lambda_{s,t,r}(\tilde{B}) |_{(\theta_s, \phi_t, \xi_r)} &= \lambda_{s,t,r}(\tilde{B}) |_{(2\pi-\theta_s, 2\pi-\phi_t, 2\pi-\xi_r)},\end{aligned}$$

we will restrict our domain to  $(0, \pi)$  instead of  $(0, 2\pi)$ . Notice that due to restrictions on  $\theta_s$ ,  $\phi_t$ , and  $\xi_r$ , and  $n$  being an even number, the end points of the interval  $(0, \pi)$  are never achieved.

For any arbitrary matrix  $K$ , we will use the notation  $\lambda_{\min}(K)$  and  $\lambda_{\max}(K)$  to denote the minimum and maximum eigenvalues respectively.

**Lemma 5.1** *If  $\theta_s$ ,  $\phi_t$ , and  $\xi_r$  lie in the interval  $(0, \pi)$ , then  $\lambda_{s,t,r}(A) > 0$ ,  $\lambda_{s,t,r}(B) > 0$ , and  $\lambda_{s,t,r}(\tilde{B}) > 0$ .*

**Proof:** We observe that  $\lambda_{\min}(A) = 4 \left( \sin^2 \frac{\theta_s}{2} + \sin^2 \frac{\phi_t}{2} + \sin^2 \frac{\xi_r}{2} \right) > 0$ . To prove other parts of the lemma, we observe that  $\lambda_{\min}(\tilde{T}) = \tilde{m} - 2\cos(\theta_s) > \frac{3+\sqrt{5}}{2} + 4\sin^2 \frac{\theta_s}{2} (= \lambda_{\min}(T)) > \frac{3+\sqrt{5}}{2}$ . Now given  $x > 1$ ,  $x + \frac{1}{x}$  increases or decreases according as  $x$  increases or decreases, thus

$$\begin{aligned}\lambda_{\min}(\tilde{P}) &= \lambda_{\min}(\tilde{T}) + \frac{1}{\lambda_{\min}(\tilde{T})} - 2 + 4\sin^2(\phi_t), \\ &> \lambda_{\min}(T) + \frac{1}{\lambda_{\min}(T)} - 2 + 4\sin^2(\phi_t), \\ &= \lambda_{\min}(P), \\ &> 1 + 4\sin^2(\phi_t) > 1.\end{aligned}$$

Consequently, we have

$$\begin{aligned}\lambda_{\min}(\tilde{B}) &= \lambda_{\min}(\tilde{P}) + \frac{1}{\lambda_{\min}(\tilde{P})} - 2 + 4\sin^2(\xi_r), \\ &> \lambda_{\min}(P) + \frac{1}{\lambda_{\min}(P)} - 2 + 4\sin^2(\xi_r), \\ &= \lambda_{\min}(B), \\ &> 4\sin^2(\xi_r) > 0.\end{aligned}$$

Notice that  $\lambda_{\min}(P) > 1$  as shown above and hence the following holds

$$\lambda_{\min}(P) + \frac{1}{\lambda_{\min}(P)} - 2 > 0.$$

Hence, the lemma is proved.

Following lemma will be useful to find a lower bound of  $\lambda_{\min}(\tilde{B}^{-1}A)$ .

**Lemma 5.2** *For sufficiently small  $h$ , we have*

$$\begin{aligned}\lambda_{\max}(\tilde{B} - A) &= \lambda_{1,1,p}(\tilde{B} - A), \\ \lambda_{\min}(\tilde{B} - A) &= \lambda_{\frac{p}{2}, \frac{p}{2}, q}(\tilde{B} - A),\end{aligned}$$

for any fixed integers  $p$  and  $q$ , satisfying  $1 \leq p, q \leq \frac{n}{2}$ .

**Proof:** We have

$$\lambda_{s,t,r}(\tilde{B} - A) = \tilde{m} - 6 + \frac{1}{\lambda_s(\tilde{T})} + \frac{1}{\lambda_{s,t}(\tilde{P})}.$$

Now,

$$\begin{aligned}\lambda_{\min}(\tilde{T}) &= \tilde{m} - 2\cos(\theta_1) = \lambda_1(\tilde{T}), \\ \lambda_{\min}(\tilde{P}) &= \lambda_{\min}(\tilde{T}) + \frac{1}{\lambda_{\min}(\tilde{T})} - 2\cos(\phi_1) = \lambda_{1,1}(\tilde{P}).\end{aligned}$$

So,

$$\begin{aligned}\lambda_{\max}(\tilde{B} - A) &= \tilde{m} - 6 + \frac{1}{\lambda_{\min}(\tilde{T})} + \frac{1}{\lambda_{\min}(\tilde{P})}, \\ &= \lambda_{1,1,p}(\tilde{B} - A).\end{aligned}$$

Also, we notice that

$$\begin{aligned}\lambda_{\max}(\tilde{T}) &= \tilde{m} - 2\cos(\theta_{n/2}) = \lambda_{n/2}(\tilde{T}), \\ \lambda_{\max}(\tilde{P}) &= \lambda_{\max}(\tilde{T}) + \frac{1}{\lambda_{\max}(\tilde{T})} - 2\cos(\phi_{n/2}) = \lambda_{\frac{n}{2},\frac{n}{2}}(\tilde{P}).\end{aligned}$$

This gives

$$\begin{aligned}\lambda_{\min}(\tilde{B} - A) &= \tilde{m} - 6 + \frac{1}{\lambda_{\max}(\tilde{T})} + \frac{1}{\lambda_{\max}(\tilde{P})}, \\ &= \lambda_{\frac{n}{2},\frac{n}{2},q}(\tilde{B} - A).\end{aligned}$$

**Corollary 5.3** *From arguments in the proof of lemma 5.1 above, we at once have*

$$\begin{aligned}\lambda_{\min}(\tilde{B}) &= \lambda_{1,1,1}(\tilde{B}), \\ \lambda_{\min}(A) &= \lambda_{1,1,1}(A).\end{aligned}$$

**Lemma 5.4** *For sufficiently small  $h$  and  $c < \frac{16}{9}\pi^2(7 + 3\sqrt{5})$ , we have*

$$\lambda_{s,t,r}(\tilde{B}) - \lambda_{s,t,r}(A) < 0,$$

*i.e.,*

$$\lambda_{\min}(\tilde{B}^{-1}A) > 1.$$

**Proof:** From previous Lemma, we have  $\lambda_{\max}(\tilde{B} - A) = \lambda_{1,1,1}(\tilde{B} - A)$ . We will prove that  $\lambda_{1,1,1}(\tilde{B} - A) < 0$ . We notice that

$$\lambda_{1,1,1}(\tilde{B} - A) = \tilde{m} - 6 + \frac{1}{\lambda_1(\tilde{T})} + \frac{1}{\lambda_{1,1}(\tilde{P})}.$$

Using truncated Taylor series expansions, we have  $\cos(\theta_1) \approx 1 - \theta_1^2 = 1 - 2\pi^2 h^2$  and  $\cos(\phi_1) \approx 1 - \phi_1^2 = 1 - 2\pi^2 h^2$ , and we have

$$\lambda_{1,1,1}(\tilde{B} - A) = \frac{(-56\pi^2 - 24\sqrt{5}\pi^2 + 10K^2)h^2 + O(h^3)}{7 + 3\sqrt{5} + O(h)},$$

$$< 0.$$

provided  $K < \frac{1}{5}\sqrt{5}\pi\sqrt{2}(3 + \sqrt{5})$ , that is,  $c < \frac{16}{9}\pi^2(7 + 3\sqrt{5})$  and thus  $\lambda_{s,t,r}(\tilde{B} - A) \leq \lambda_{1,1,1}(\tilde{B} - A) < 0$ , and from this it follows that  $\lambda_{\min}(\tilde{B}^{-1}A) > 1$ .

Hence, the lemma is proved.

Next we will bound  $\lambda_{\max}(\tilde{B}^{-1}A)$ , for which we will use the following lemma.

**Lemma 5.5** *For sufficiently small  $h$ , we have*

$$\frac{1}{(6 - \tilde{m})}(\lambda_{s,t,r}(A) - \lambda_{s,t,r}(\tilde{B})) \leq \frac{1}{(6 - m)}(\lambda_{s,t,r}(A) - \lambda_{s,t,r}(B)).$$

**Proof:** We shall treat variables  $\theta_s$  and  $\phi_t$  as continuous variables  $\theta$  and  $\phi$  respectively. Also, we assume that  $h$  is any continuous variable, not necessarily the mesh size. We define

$$f(\theta, \phi, h) = \frac{1}{(6 - \tilde{m})}(\lambda_{s,t,r}(A) - \lambda_{s,t,r}(\tilde{B})).$$

Now taking partial derivative of  $f(\theta, \phi, h)$  with respect to  $h$ , we obtain

$$\frac{\partial}{\partial h} f(\theta, \phi, h) = \frac{f(\theta, \phi)h^5 + g(\theta, \phi)h^4 + h(\theta, \phi)h^3 + u(\theta, \phi)h^2 + v(\theta, \phi)h^1 + w(\theta, \phi)}{q(\theta, \phi, h)},$$

where

$$q(\theta, \phi, h) = (58 + 14\sqrt{5} - 2y\sqrt{5} - 28x + 4x^2 - 4\sqrt{5}x + 4yx - 14y)^2 (7 + \sqrt{5} - 2x)^2$$

$$\times (-5 + \sqrt{5})^2 + O(h).$$

Here  $x = 2\cos(\theta)$  and  $y = 2\sin(\theta)$ , and

$$\begin{aligned} f(\theta, \phi) &= -1024 K^6, \\ g(\theta, \phi) &= -16 K(-288 K^4 x - 80 K^4 y + 160 K^4 \sqrt{5} + 928 K^4), \\ h(\theta, \phi) &= -16 K(6016 K^3 + 1856 \sqrt{5} K^3 - 576 \sqrt{5} K^3 x - 160 \sqrt{5} K^3 y + 288 K^3 yx \\ &\quad + 512 K^3 x^2 + 32 y^2 K^3 - 928 K^3 y - 3264 K^3 x), \\ u(\theta, \phi) &= -16 K(8224 \sqrt{5} K^2 - 4896 \sqrt{5} K^2 x - 80 y^2 K^2 x - 4528 K^2 y + 20864 K^2 \\ &\quad + 2448 K^2 yx - 15440 K^2 x + 4224 K^2 x^2 + 768 \sqrt{5} K^2 x^2 + 240 y^2 K^2 - 448 K^2 x^3 \\ &\quad - 384 K^2 yx^2 - 1392 \sqrt{5} K^2 y + 48 y^2 \sqrt{5} K^2 + 432 \sqrt{5} K^2 yx), \\ v(\theta, \phi) &= -16 K(4224 \sqrt{5} K x^2 - 2368 K x^3 - 34448 K x + 192 K x^4 - 4128 \sqrt{5} y K \\ &\quad + 12832 K x^2 - 14000 \sqrt{5} K x - 10400 y K + 7728 K y x + 16224 \sqrt{5} K - 368 y^2 K x \\ &\quad - 80 y^2 \sqrt{5} K x + 624 y^2 K - 384 K x^2 y \sqrt{5} - 448 \sqrt{5} K x^3 + 224 K x^3 y \\ &\quad - 2112 K x^2 y + 36864 K + 2448 \sqrt{5} K y x + 64 y^2 K x^2 + 240 y^2 \sqrt{5} K), \\ w(\theta, \phi) &= -16 K(25216 + 11712 \sqrt{5} - 8920 y - 28840 x + 13584 x^2 + 8496 y x \\ &\quad - 3408 x^3 + 480 x^4 - 3200 y x^2 + 592 x^3 y + 496 y^2 - 408 y^2 x + 128 y^2 x^2 \\ &\quad - 32 x^5 - 48 x^4 y - 16 y^2 x^3 - 4040 y \sqrt{5} - 13144 \sqrt{5} x + 5776 \sqrt{5} x^2 \\ &\quad - 1184 \sqrt{5} x^3 - 184 y^2 \sqrt{5} x + 3504 \sqrt{5} y x + 272 y^2 \sqrt{5} - 1056 \sqrt{5} x^2 y \\ &\quad + 112 \sqrt{5} x^3 y + 32 y^2 \sqrt{5} x^2 + 96 x^4 \sqrt{5}). \end{aligned}$$

With the aid of symbolic manipulator *MAPLE* [12], we find that

$$\begin{aligned}
q(\theta, \phi, h) &> 0, \\
f(\theta, \phi) &= (-1024) K^6, \\
g(\theta, \phi) &< (-3072 - 2560\sqrt{5}) K^5, \\
h(\theta, \phi) &< (-15360 - 6144\sqrt{5}) K^4, \\
u(\theta, \phi) &< (-10240\sqrt{5} - 20480) K^3, \\
v(\theta, \phi) &< (-5120\sqrt{5} - 12800) K^2, \\
w(\theta, \phi) &< 0.
\end{aligned}$$

Note that  $K = \frac{3\sqrt{5}\sqrt{c}}{10} > 0$  and from this it follows that  $\frac{\partial}{\partial h} f(\theta, \phi, h) < 0$ , and this indicates that  $f(\theta, \phi, h)$  is a decreasing function, or in other words, the function  $f(\theta, \phi, h)$  increases as  $h$  tends to zero and it tends to  $f(\theta, \phi, 0)$ . With this we have established the lemma.

**Lemma 5.6** *For MNF preconditioned isotropic operator, we have*

$$\lambda_{max}(\tilde{B}^{-1}A) \leq \frac{1}{K}h^{-1},$$

where  $K = \frac{3\sqrt{5}\sqrt{c}}{10}$ .

**Proof:** We have

$$\begin{aligned}
\lambda_{s,t,r}(\tilde{B}^{-1}A) &= \frac{\lambda_{s,t,r}(A)}{\lambda_{s,t,r}(\tilde{B})} = \frac{1}{1 - \frac{\lambda_{s,t,r}(A) - \lambda_{s,t,r}(\tilde{B})}{\lambda_{s,t,r}(A)}}, \\
&= \frac{1}{1 - (6 - \tilde{m}) \frac{1 - \frac{1}{(6-\tilde{m})\lambda_s(\tilde{T})} - \frac{1}{(6-\tilde{m})\lambda_{s,t}(\tilde{P})}}{\lambda_{s,t,r}(A)}}.
\end{aligned}$$

Using  $\tilde{m} \approx \frac{7+\sqrt{5}}{2} + Kh$ , we have

$$\lambda_{s,t,r}(\tilde{B}^{-1}A) = \frac{1}{1 - (\frac{5-\sqrt{5}}{2} - Kh) \frac{1 - \frac{1}{(6-\tilde{m})\lambda_s(\tilde{T})} - \frac{1}{(6-\tilde{m})\lambda_{s,t}(\tilde{P})}}{\lambda_{s,t,r}(A)}}.$$

We observe that  $-\lambda_{s,t,r}(B) \leq 0$  (Lemma 5.1), and it follows that

$$\begin{aligned}
\frac{\lambda_{s,t,r}(A) - \lambda_{s,t,r}(B)}{\lambda_{s,t,r}(A)} &\leq 1, \\
\Leftrightarrow (6-m) \frac{\left(1 - \frac{1}{(6-m)\lambda_s(\tilde{T})} - \frac{1}{(6-m)\lambda_{s,t}(\tilde{P})}\right)}{\lambda_{s,t,r}(A)} &\leq \frac{2}{5-\sqrt{5}}(6-m), \\
\Leftrightarrow \frac{\left(1 - \frac{1}{(6-m)\lambda_s(\tilde{T})} - \frac{1}{(6-m)\lambda_{s,t}(\tilde{P})}\right)}{\lambda_{s,t,r}(A)} &\leq \frac{2}{5-\sqrt{5}}.
\end{aligned}$$

Using lemma 5.5, the theorem is proved.

Finally, we prove theorem (3.3).

**Proof of Theorem (3.3):** We have

$$\kappa(\tilde{B}^{-1}A) = \frac{\lambda_{max}(\tilde{B}^{-1}A)}{\lambda_{min}(\tilde{B}^{-1}A)}.$$

From lemma (5.4), we have  $\lambda_{min}(\tilde{B}^{-1}A) > 1$ . We prove that this bound is tight, i.e.,  $\lambda_{min}(\tilde{B}^{-1}A) = O(1)$ . Choosing  $(s, t, r) = (1, 1, 1)$ , we have  $\lambda_{1,1,1}(A) \approx 12\pi^2 h^2$  and

$$\begin{aligned} \lambda_{1,1,1}(\tilde{B}) &\approx -5/2 + 1/2\sqrt{5} + Kh + 12\pi^2 h^2 + (3/2 + 1/2\sqrt{5} + Kh + 4\pi^2 h^2)^{-1} \\ &\quad + (-1/2 + 1/2\sqrt{5} + Kh + 8\pi^2 h^2 + (3/2 + 1/2\sqrt{5} + Kh + 4\pi^2 h^2)^{-1})^{-1}. \end{aligned}$$

After algebraic manipulations and collecting powers of  $h$ , we have

$$\lambda_{1,1,1}(\tilde{B}^{-1}A) = \frac{\lambda_{1,1,1}(A)}{\lambda_{1,1,1}(\tilde{B})} \approx \frac{3\pi^2(3 + \sqrt{5})^2}{14\pi^2 + 5K^2 + 6\sqrt{5}\pi^2} = O(1),$$

so that we have  $\lambda_{min}(\tilde{B}^{-1}A) = O(1)$ . Next, we wish to prove that  $\lambda_{max}(\tilde{B}^{-1}A) = O(h^{-1})$ . For this we fix  $s = \frac{1}{\sqrt{2\pi h}}$ ,  $t = 1$ , and  $r = 1$ . Now, using truncated Taylor series expansion and ignoring higher powers of  $h$ , we have,  $\cos(\theta_{\frac{1}{\sqrt{2\pi h}}}) \approx 1 - \pi h$ ,  $\cos(\phi_1) \approx 1 - 2\pi^2 h^2$ , and  $\cos(\xi_1) \approx 1 - 2\pi^2 h^2$  and hence

$$\begin{aligned} \lambda_{\frac{1}{\sqrt{2\pi h}}, 1, 1}(\tilde{B}^{-1}A) &\approx \frac{\pi(3 + \sqrt{5})^2 h}{(12\pi^2\sqrt{5} + 68\pi^2 + 40K\pi + 10K^2)h^2}, \quad (18) \\ &= O(h^{-1}). \end{aligned}$$

This gives

$$\begin{aligned} \lambda_{max}(\tilde{B}^{-1}A) &= O(h^{-1}), \\ \Rightarrow \kappa(\tilde{B}^{-1}A) &= O(h^{-1}). \end{aligned}$$

Hence the theorem.

**Theorem 5.7** For Nested Factorization preconditioned isotropic operator, we have

$$\kappa(B^{-1}A) \geq O(h^{-1}).$$

**Proof:** Choosing  $s = t = r = 1$ , we have

$$\lambda_{min}(B^{-1}A) \leq \lambda_{1,1,1}(B^{-1}A) = 24 \frac{\pi^2(3 + \sqrt{5})(6 + 2\sqrt{5}) + O(h^2)}{224\pi^2 + 96\sqrt{5}\pi^2 + O(h^2)},$$

and choosing  $s = \frac{1}{\sqrt{2\pi h}}$  and  $t = r = 1$ , we have

$$\lambda_{max}(B^{-1}A) \geq \lambda_{\frac{1}{\sqrt{2\pi h}}, 1, 1}(B^{-1}A) = 4 \frac{\pi (3 + \sqrt{5}) (6 + 2\sqrt{5}) + O(h)}{h (544 \pi^2 + 96 \pi^2 \sqrt{5}) + O(h^2)},$$

thus,  $\kappa(B^{-1}A) \geq \frac{\lambda_{\frac{1}{\sqrt{2\pi h}}, 1, 1}(B^{-1}A)}{\lambda_{1, 1, 1}(B^{-1}A)}$ , i.e.,

$$\kappa(B^{-1}A) \geq O(h^{-1}).$$

Hence, the lemma is proved.

**Result:** For the MNF preconditioned isotropic operator, the optimal value of the parameter  $c$  occurs near  $c_p = \left(\frac{32}{3} - \frac{8}{9} \sqrt{5} \sqrt{19 + 6\sqrt{5}} + 8/3 \sqrt{5}\right) \pi^2$  for the periodic problem, and  $\frac{1}{4}c_p$  for the corresponding Dirichlet problem.

In principle, we would like to choose  $c$  to minimize the condition number  $\kappa(\tilde{B}^{-1}A)$ . Unfortunately, this turns out to involve rather complicated calculations. Instead, we use the same expression for the condition number where we obtain the exact order of condition number in Theorem (5.6). We have  $\kappa(\tilde{B}^{-1}A) = O(h^{-1})$  and this order of condition number is achieved for  $\lambda_{max}(\tilde{B}^{-1}A)$  corresponding to the grid point  $\left(\frac{1}{\sqrt{2\pi h}}, 1, 1\right)$  and for  $\lambda_{min}(\tilde{B}^{-1}A)$  corresponding to the grid point  $(1, 1, 1)$ . We obtain

$$\begin{aligned} \kappa(\tilde{B}^{-1}A) &= \frac{\lambda_{\frac{1}{\sqrt{2\pi h}}, 1, 1}(\tilde{B}^{-1}A)}{\lambda_{1, 1, 1}(\tilde{B}^{-1}A)}, \\ &\approx 1/3 \frac{14 \pi^2 + 5 K^2 + 6 \sqrt{5} \pi^2}{\pi (12 \sqrt{5} \pi^2 + 68 \pi^2 + 40 K \pi + 10 K^2) h}, \\ \frac{\partial}{\partial K} \kappa(\tilde{B}^{-1}A) &\approx 10/3 \frac{-14 \pi^2 + 10 K \pi + 5 K^2 - 6 \sqrt{5} \pi^2}{(6 \sqrt{5} \pi^2 + 34 \pi^2 + 20 K \pi + 5 K^2)^2 h}. \end{aligned}$$

To find optimal value of parameter  $c$ , we set  $\frac{\partial}{\partial K} \kappa(\tilde{B}^{-1}A) = 0$ . Now choosing positive root of  $K$ , we get

$$K = -\pi + 1/5 \sqrt{95 \pi^2 + 30 \sqrt{5} \pi^2}.$$

Consequently, we have

$$c_p = 5.8\pi^2 \approx 52.43.$$

On the other hand, to predict the optimal value of the parameter  $c_d$ , for the Dirichlet problem, we use  $\delta_p = c_p h^2 = \delta_d = c_d h_d^2$  as in [7, 9], where  $h = \frac{h_d}{2}$  and we obtain

$$c_d = \frac{1}{4}c_p \approx 1.45\pi^2 \approx 13.10. \quad (19)$$

We note here that the optimal value of parameter for the Periodic and Dirichlet problems for MILU are  $c_p \approx 12\pi^2$  and  $c_d \approx 3\pi^2$  respectively [9].

## References

- [1] Appleyard JR, and Cheshire IM. Nested Factorization. *SPE 12264, presented at the Seventh SPE Symposium on Reservoir Simulation, San Francisco, USA* 1983;
- [2] Saad Y. *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, Boston, MA, 1996.
- [3] Kuznetsova NN, Diyankov OV, Kotegov SS, Koshelev SV, Krasnogorov IV, Pravilnikov VY, and Maliassov SY. The family of nested factorizations. *Russian Journal of Numerical Analysis and Mathematical Modelling* 2007; **22**(4): 393-412.
- [4] Gustafsson I. A class of first order factorization methods. *BIT* 1978; **18**(2): 142-156.
- [5] Appleyard JR. Proof that Simple Colsum modified ILU Factorizations of Matrices arising in Fluid Flow Problems are not Singular. Available online at the link : <http://www.polyhedron.com/colsum-proof>.
- [6] Crumpton PA, Fjerstad PA, and Berge J. Parallel computing, Using ECLIPSE Parallel. *White Paper* 2003 Available online at the link : [www.sis.slb.com/media/about/whitepaper\\_parallelcomputing.pdf](http://www.sis.slb.com/media/about/whitepaper_parallelcomputing.pdf).
- [7] Chan TF, and Elman HC. Fourier analysis of iterative methods for elliptic boundary value problems. *SIAM Review* 1989; **31**(1): 20-49.
- [8] Otto K. Analysis of Preconditioners for Hyperbolic Partial Differential Equations. *SIAM J. Num. Ana.* 1996; **33**(6): 2131-2165.
- [9] Donato JM, and Chan TF. Fourier analysis of incomplete factorization preconditioners for three dimensional anisotropic problems. *SIAM J. Sci. Stat. Comput.* 1992; **13**(1): 319-338.
- [10] Davis PJ. *Circulant Matrices*. Wiley. 1979.
- [11] Bai Z, Demmel J, Dongarra J, Ruhe A, and van der Vorst H. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. SIAM, Philadelphia, 2000.
- [12] Char BW, Fee GJ, Geddes KO, Gonnet GH, and Monagan MB. A tutorial introduction to Maple. *J.Symb. Comput.* 1986; **2**(2):179-200.
- [13] Made MMM, and Notay Y. Dynamically relaxed block incomplete factorizations for solving two- and three-dimensional problems. *SIAM J. Sci. Stat. Comput.* 2000; **21**(6): 2008-2028.
- [14] Meurant G. *Computer Solution of Large Linear Systems* North-Holland Publishing Co., Amsterdam, 1999.



---

Centre de recherche INRIA Saclay – Île-de-France  
Parc Orsay Université - ZAC des Vignes  
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex  
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier  
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq  
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique  
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex  
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex  
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex  
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

---

Éditeur  
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)  
<http://www.inria.fr>  
ISSN 0249-6399