Almost Optimal Order Approximate Inverse Based Preconditioners For 3-d Convection Dominated Problems On Tensor-grids

dr. J.M. Maubach

July 17, 2006

Abstract
For a one-dimensional diffusion problem on an refined computational grid we present preconditioners based on the standard approximate inverse technique. Next, we determine its spectral condition number κ_2 and perform numerical calculations which corroborate the result. Then we perform numerical calculations which show that the standard approximate inverse preconditioners and our modified versions behave in a similar manner. To finish with we show that a combination of the standard approximate inverse with an additional incomplete factorization leads to an *almost optimal order* preconditioner in 1, 2 and 3 dimensions, with or without dominant convection.

1 Introduction

Let $\mathbf{A} \in \mathcal{R}^{n\times n}$ and $\mathbf{b} \in \mathcal{R}^n$ be a coefficient matrix and the right hand side vector and assume that $\mathbf{x} \in \mathcal{R}^n$ is the solution of $\mathbf{Ax} = \mathbf{b}$. The amount of iterations for the solution of the system $\mathbf{Ax} = \mathbf{b}$ depends on the spectral condition number $\kappa_2(\mathbf{A})$ (see [14] and [13] for a one step fixed point, steepest descent, conjugate gradient, and GMRES method).

To reduce the amount of iterations one solves a preconditioned (transformed) system $\hat{\mathbf{A}}\hat{\mathbf{x}} = \hat{\mathbf{b}}$ for which $\kappa_2(\hat{\mathbf{A}}) << \kappa_2(\mathbf{A})$, and afterwards determines $\mathbf{x}$ from $\hat{\mathbf{x}}$. The optimal case is when $\kappa_2(\hat{\mathbf{A}}) = O(1)$ for $n \to \infty$. For elliptic problems without convection such optimal transformations exist (see for instance [2], [3]), but not for problems with (more general) convection.

The problem of interest is a convection diffusion partial differential equation, discretized with a Finite Difference Method on a non-uniform computational mesh. The mesh zooms in into a point: The distance between subsequent points grows with a fixed factor ψ, in practice often $1/2$, see [12], [11] and [9], independent on the dimension of the problem. We are interested in the performance of the standard approximate inverse preconditioner: See [8], [16], [4] and [6].

The motivation is that this preconditioner is black-box with respect to the presence or absence of convection and that we can present optimal numerical results in the convection dominated case.

The to be presented spectral condition number derivation is (1) restricted to one spatial dimension, (2) sole diffusion and our (3) modified version of the standard approximate inverse. One dimension because we need (1) [10] to calculate an explicit matrix inverse and (2) just diffusion and (3) modification from the standard because we use a Toeplitz matrix approach from [5]. The results are based on the PhD thesis [7] and have not been published elsewhere.

2 The convection diffusion problem

Let $\Omega \subset \mathcal{R}^d$ be smooth and let $\mathbf{v} \in \mathcal{R}^d$. The problem of interest is the convection diffusion equation:

$$\begin{aligned} -\triangle u + \mathbf{v}\nabla u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{at } \partial\Omega. \end{aligned} \tag{1}$$

The partial differential equation is discretized with a Finite Difference Method on a tensor-grid (examples are Figures 1 and 2).

To be able to apply the tensor-grids we choose as $\Omega \subset \mathcal{R}^d$ the unit d-cube. The gridpoints are of the form $(x_i^{(1)}, x_j^{(2)}, x_k^{(3)})$. For $k = 1, 2$ and 3 and related integer $n_k > 0$ let $f_k := 1/\psi^{n_k+1}$,

$$F_k \colon x \mapsto \frac{f_k^{(1-x)} - f_k}{1 - f_k} \tag{2}$$

and define

$$x_i^{(k)} = F_k(i/(n_k+1)) = \frac{f_k^{1-i/(n_k+1)} - f_k}{1 - f_k}, \qquad i = 0, \dots, n_k + 1. \tag{3}$$

Due to this construction $x_0 = 0$, $x_{n+1} = 1$ and

$$\begin{aligned} \frac{x_{i+1} - x_i}{x_i - x_{i-1}} &= \frac{f^{1-(i+1)/(n+1)} - f^{1-i/(n+1)}}{f^{1-i/(n+1)} - f^{1-(i-1)/(n+1)}} = \frac{f^{1-1/(n+1)} - f}{f - f^{1+1/(n+1)}} \\ &= \frac{f^{-1/(n+1)} - 1}{1 - f^{1/(n+1)}} = \frac{\psi - 1}{1 - \psi^{-1}} \\ &= \frac{\psi}{\psi} \frac{\psi - 1}{1 - \psi^{-1}} = \psi \frac{\psi - 1}{\psi - 1} = \psi. \end{aligned} \tag{4}$$

The amount of interior gridpoints in the x_k-direction is n_k and the amount of degrees of freedom is $n = \Pi_{k=1}^d n_k$.

For the sake of presentation we focus for the case $d = 1$ and we assume that $\mathbf{v} = (v_1, v_2, v_3)$ where v_i are all non-negative. The definition of the distances:

$$h_i := x_{i+1} - x_i, \qquad \forall i = 0, \dots, n, \tag{5}$$

in combination with (4) implies that

$$h_i = \psi^i H, \qquad \forall i = 0, \ldots, n$$

where

$$1 - 0 = \sum_{i=0}^{n}(x_{i+1} - x_i) = \sum_{i=0}^{n} h_i = \sum_{i=0}^{n} \psi^i H = H \cdot \frac{1-\psi^{n+1}}{1-\psi} \implies H = \frac{1-\psi}{1-\psi^{n+1}}.$$

With the use of the grid points x_i and their distances h_i we construct our FDM approximation. The second derivatives are approximated with a second order finite difference at x_i:

$$\frac{2}{h_{i-1}+h_i}\left(\frac{u_{i+1}-u_i}{h_i} - \frac{u_i - u_{i-1}}{h_{i-1}}\right) = \frac{1}{h_i^2} \cdot \frac{2\psi}{1+\psi}\left(-\psi u_{i-1} + (1+\psi)u_i - u_{i+1}\right)$$

and the first derivatives are approximated with a first order backward finite difference at x_i:

$$\frac{1}{h_i}\left(-u_{i-1} + u_i\right).$$

In stencil notation, the finite difference operator for (1) is

$$\frac{1}{h_i^2} \cdot \frac{2\psi}{1+\psi}\left[-\psi, (1+\psi), -1\right] + v \cdot \frac{1}{h_i}\left[-1, 1, 0\right], \qquad i = 1, \ldots, n. \tag{6}$$

After the elimination of the Dirichlet degrees of freedom u_0 and u_n one obtains a linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ where $A \in \mathcal{R}^{n \times n}$ and $\mathbf{x}, \mathbf{b} \in \mathcal{R}^n$ where $\mathbf{A}$ is positive definite by construction.

3 The preconditioned system and its Toeplitz approximation

Given a certain sparsity pattern, the application of the approximate inverse technique calculates a matrix $\mathbf{G}_n$ from $\mathbf{A}_n$ such that $\mathbf{A}_n\mathbf{G}_n$ approximates $\mathbf{I}_n$. For our convection diffusion problem on our tensor-grid, we use the standard sparsity pattern

$$\mathcal{S} = \{(i,j) \colon [\mathbf{A}_n]_{ij} \neq 0\}.$$

and standard Frobenius norm minimization. Let $\mathbf{g}_i^{(n)}$ be the i-th column $\mathbf{G}_n$. Due to its definition the column $\mathbf{g}_i^{(n)}$ approximates $\mathbf{A}_n^{-1}\mathbf{e}_i$. In the 1-dimensional case:

$$\mathbf{g}_i^{(n)} = [0, \ldots, 0, \underbrace{\begin{bmatrix} c + \frac{2a\psi^3}{h_i^2} + \frac{b\psi}{h_i} & \frac{-2a\psi^3}{(1+\psi)h_i^2} & 0 \\ \frac{-2a\psi^2}{(1+\psi)h_i^2} - \frac{b}{h_i} & c + \frac{2a\psi}{h_i^2} + \frac{b}{h_i} & \frac{-2a\psi}{(1+\psi)h_i^2} \\ 0 & \frac{-2a}{(1+\psi)h_i^2} - \frac{b}{\psi h_i} & c + \frac{2a}{\psi h_i^2} + \frac{b}{\psi h_i} \end{bmatrix}^{-1}}_{\mathbf{A}_{(i)}} \mathbf{e}_2, 0, \ldots, 0],$$

$$\mathbf{g}_1^{(n)} = [\begin{bmatrix} c + \frac{2a\psi}{h_1^2} + \frac{b}{h_1} & \frac{-2a\psi}{(1+\psi)h_1^2} \\ \frac{-2a}{(1+\psi)h_1^2} - \frac{b}{\psi h_1} & c + \frac{2a}{\psi h_1^2} + \frac{b}{\psi h_1} \end{bmatrix}^{-1} \mathbf{e}_1, 0, \ldots, 0],$$

$$\mathbf{g}_n^{(n)} = [0, \ldots, 0, \begin{bmatrix} c + \frac{2a\psi^3}{h_n^2} + \frac{b\psi}{h_n} & \frac{-2a\psi^3}{(1+\psi)h_n^2} \\ \frac{-2a\psi^2}{(1+\psi)h_n^2} - \frac{b}{h_n} & c + \frac{2a\psi}{h_n^2} + \frac{b}{h_n} \end{bmatrix}^{-1} \mathbf{e}_2].$$

for $i = 2, \ldots, n-1$. The three non-zero coefficients of $\mathbf{g}_i^{(n)}$ are:

$$\frac{2\,a\,\psi^2\,(2\,a + h_i\,(b + c\,\psi\,h_i))}{(1+\psi)\,h_i^4},$$
$$\frac{(2\,a\,\psi^3 + h_i\,(b\,\psi + c\,h_i))\,(2\,a + h_i\,(b + c\,\psi\,h_i))}{\psi\,h_i^4},$$
$$\frac{(2\,a\,\psi + b\,(1+\psi)\,h_i)\,(2\,a\,\psi^3 + h_i\,(b\,\psi + c\,h_i))}{\psi\,(1+\psi)\,h_i^4},$$

divided by $\det(\mathbf{A}_{(i)})$. Below we omit the superscript n of $\mathbf{g}_i^{(n)}$.

In the remainder of this paper we use the properties of a *Toeplitz matrix* because for such matrices the spectral condition number can be computed with the use of the $\|\cdot\|_\infty$ norm. We first summarize some to be used properties from [5].

Definition 1. Let $\mathbf{b} := (\ldots, b_{-2}, b_{-1}, b_0, b_1, b_2, \ldots)^T \in \mathcal{R}^\infty$, a real vector in l^1. The infinite matrix $\mathbf{B}_n$ defined by

$$\mathbf{B}_n = \begin{bmatrix} b_0 & b_1 & b_2 & \ldots \\ b_{-1} & b_0 & b_1 & \ldots \\ b_{-2} & b_1 & b_0 & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

is called a *Toeplitz matrix*. The numbers $\ldots, b_{-2}, b_{-1}, b_0, b_1, b_2, \ldots$ do *not* depend on n.

A finite section of a Toeplitz matrix, i.e. the matrix (see [5, (2.5)])

$$\mathbf{B}_n = \begin{bmatrix} b_0 & b_1 & \ldots & b_{n-1} \\ b_{-1} & b_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & b_1 \\ b_{-n+1} & \ldots & b_{-1} & b_0 \end{bmatrix}, \qquad b_i \in \mathcal{R},$$

is also called a Toeplitz matrix. For Toeplitz matrices the spectral condition number can be computed using the $\|\cdot\|_\infty$ norm. This is due to a special relation between the largest and smallest singular values $\sigma_1(\mathbf{B}_n)$ and $\sigma_n(\mathbf{B}_n)$ of a matrix $\mathbf{B}_n$ and the values of $\|\mathbf{B}_n\|_\infty$ and $\|\mathbf{B}_n^{-1}\|_\infty$. For the largest singular value we have (see [5, (2.5) and Theorem 4.13]):

Theorem 1. *For* $\mathbf{b} \in l^1$ *the following holds*

$$\lim_{n\to\infty} \sigma_n(\mathbf{B}_n) = \lim_{n\to\infty} \|\mathbf{B}_n\|_\infty = \sum_{i=-\infty}^{\infty} |b_i|. \tag{7}$$

There is a similar result for the smallest singular value σ_1 (see [5, Theorem 4.3 (4.8)]):

Theorem 2. *If $\sigma_1(\mathbf{B}_n)$ is such that for all $n \in \mathcal{N}$, $\sigma_1(\mathbf{B}_n) \geq \delta > 0$ then*

$$\lim_{n\to\infty} \sigma_1(\mathbf{B}_n) = \lim_{n\to\infty} \frac{1}{\|\mathbf{B}_n^{-1}\|_\infty}. \tag{8}$$

Therefore, if $\mathbf{B}_n$ is Toeplitz, we have

$$\lim_{n\to\infty} \kappa_2(\mathbf{B}_n) = \lim_{n\to\infty} \frac{\sigma_n(\mathbf{B}_n)}{\sigma_1(\mathbf{B}_n)} = \frac{\lim_{n\to\infty}\sigma_n(\mathbf{B}_n)}{\lim_{n\to\infty}\sigma_1(\mathbf{B}_n)} = \sum_{i=-\infty}^{\infty} |b_i| / \lim_{n\to\infty} \|\mathbf{B}_n^{-1}\|_\infty.$$

Formulas for eigenvalues and eigenvectors of a tridiagonal Toeplitz matrix:

Theorem 3. *A tridiagonal Toeplitz matrix $\mathbf{B}_n$ $a_{-1}a_1 \neq 0$ has eigenvalues:*

$$\sigma(\mathbf{B}_n) = \left(a_0 + 2\sqrt{a_{-1}a_1}\cos(\frac{p\pi}{n+1})\right)_{p=1}^{n} \tag{9}$$

and related eigenvectors $\mathbf{x}_p$:

$$\mathbf{x}_p = \left(\sqrt{(a_1/a_{-1})}^{\,i} \sin(\frac{p\pi}{n+1} i)\right)_{i=0}^{n}.$$

In order to calculate the spectral condition number of $\mathbf{G}_n\mathbf{A}_n$ and $\mathbf{A}_n\mathbf{G}_n$ we need to determine the explicit form of these products. We first focus on $\mathbf{A}_n\mathbf{G}_n$, a five diagonal matrix of which diagonals -1 and $+1$ contain zeros due to its the construction. The lower and upper codiagonals of $\mathbf{A}_n\mathbf{G}_n$ can be calculated in a straightforward manner:

$$\begin{bmatrix} \ddots & \ddots & & & & & \\ \ddots & a_{i-2,i-2} & a_{i-2,i-1} & 0 & 0 & 0 & \\ & a_{i-1,i-2} & a_{i-1,i-1} & a_{i-1,i} & 0 & 0 & \\ & 0 & a_{i,i-1} & a_{i,i} & a_{i,i+1} & 0 & \\ & 0 & 0 & a_{i+1,i} & a_{i+1,i+1} & a_{i+1,i+2} & \\ & 0 & 0 & 0 & a_{i+2,i+1} & a_{i+2,i+2} & \ddots \\ & & & & & \ddots & \ddots \end{bmatrix} \cdot \begin{bmatrix} \vdots \\ 0 \\ \mathbf{g}_{(i),i-1} \\ \mathbf{g}_{(i),i} \\ \mathbf{g}_{(i),i+1} \\ 0 \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ a_{i-2,i-1}\cdot\mathbf{g}_{(i),i-1} \\ 0 \\ 1 \\ 0 \\ a_{i+2,i+1}\cdot\mathbf{g}_{(i),i+1} \\ \vdots \end{bmatrix}.$$

implies that

$$a_{i-2,i-1}\cdot\mathbf{g}_{(i),i-1} = \frac{-2\psi^5}{(1+\psi)\ h_i^2}\mathbf{g}_{(i),i-1} = \frac{-2\psi^5}{(1+\psi)\ h_i^2}\frac{2\psi^2\ (2+h_i v)}{(1+\psi)\,h_i^4}/\det(\mathbf{A}_{(i)})$$

and likewise for the lower co-diagonal $a_{i+2,i+1}\cdot\mathbf{g}_{(i),i+1}$. Without convection ($v = 0$):

$$\det(\mathbf{A}_{(i)}) = \frac{8a^3\psi^3(1+\psi^2)}{h_i^6(1+\psi)^2}, \qquad i = 2,\ldots,n-1,$$

and

$$\mathbf{g}_{(i)} = h_i^2 \cdot \frac{1+\psi}{2a(1+\psi^2)} \cdot [0, \dots, 0, 1/\psi, (1+\psi)/\psi, 1, 0, \dots, 0], \qquad i = 2, \dots, n-1. \quad (10)$$

This leads to upper co-diagonal

$$a_{i-2,i-1} \cdot \mathbf{g}_{(i),i-1} = \frac{-2\psi^5}{(1+\psi)\ h_i^2}\mathbf{g}_{(i),i-1} = \frac{-2\psi^5}{(1+\psi)\ h_i^2} \cdot h_i^2 \cdot \frac{1+\psi}{2(1+\psi^2)} \cdot 1/\psi = -\psi^4/(1+\psi^2),$$

and lower co-diagonal

$$a_{i+2,i+1} \cdot \mathbf{g}_{(i),i+1} = \frac{-2}{(1+\psi)\,\psi^2\, h_i^2} \cdot h_i^2 \cdot \frac{1+\psi}{2(1+\psi^2)} = -\frac{1}{\psi^2(1+\psi^2)}.$$

Note that these representations are independent of h_i and even a possible diffusion factor $a \neq 1$ would drop out.

Even in the absence of convection, the matrix $\mathbf{A}_n\mathbf{G}_n$ is not Toeplitz because just a few of its entries spoil this property. For the numerical analysis in the remainder we therefore focus on the case of no convection and on the obvious Toeplitz approximations $\mathbf{B}_n^{\mathrm{r}}$ of $\mathbf{A}_n\mathbf{G}_n$ and $\mathbf{B}_n^{\mathrm{l}}$ of $\mathbf{G}_n\mathbf{A}_n$:

$$\mathbf{B}_n^{\mathrm{r}} = [-\frac{1}{\psi^2(1+\psi^2)}, 0, 1, 0, -\frac{\psi^4}{1+\psi^2}], \qquad \mathbf{B}_n^{\mathrm{l}} = [-\frac{\psi^2}{1+\psi^2}, 0, 1, 0, -\frac{1}{1+\psi^2}]. \quad (11)$$

Note that both approximations $\mathbf{B}_n^{\mathrm{r}}$ and $\mathbf{B}_n^{\mathrm{l}}$ are reducible: Due to its special diagonal structure, the degrees of freedom can be permuted with a permutation $\mathbf{P}$ (first all odd numbers, then all even numbers) such that

$$\bar{\mathbf{B}}_n^{\mathrm{l}} = \mathbf{P}\mathbf{B}_n^{\mathrm{l}}\mathbf{P}^T = \begin{bmatrix} \mathbf{B}_{n/2}^{\mathrm{l}} & \\ & \mathbf{B}_{n/2}^{\mathrm{l}} \end{bmatrix} \quad (12)$$

Because eigenvalues and the spectral condition number are permutation invariant the spectrum of $\bar{\mathbf{B}}_n^{\mathrm{l}}$ and $\mathbf{B}_n^{\mathrm{l}}$ and $\mathbf{B}_{n/2}^{\mathrm{l}}$ are identical.

Example 1. *Let $c = 0$ and $\psi = 1/2$. Then i-th finite difference equation* (6) *is:*

$$\begin{aligned} \frac{1}{h_i^2} \cdot \frac{2\psi}{1+\psi}[-\psi, (1+\psi), -1] &= \frac{1}{H^2}\frac{1}{(\psi^2)^i} \cdot \frac{2\psi}{1+\psi}[-\psi, (1+\psi), -1] \\ &= \frac{1}{H^2}4^i \cdot \frac{2}{3}\left[-\frac{1}{2}, \frac{3}{2}, -1\right] \\ &= \frac{1}{H^2}4^i \cdot \left[-\frac{1}{3}, 1, -\frac{2}{3}\right] \end{aligned}$$

for all $i = 1, \ldots, n$. *This leads to*

$$\mathbf{A}_n = \frac{4}{H^2} \cdot \begin{bmatrix} 1 & -\frac{2}{3} & & & & \\ -\frac{4}{3} & 4 & -\frac{8}{3} & & & \\ & -\frac{16}{3} & 16 & -\frac{32}{3} & & \\ & & -\frac{64}{3} & 64 & \ddots & \\ & & & \ddots & \ddots & -\frac{2}{3} \cdot 4^{n-2} \\ & & & & -\frac{1}{3} \cdot 4^{n-1} & 4^{n-1} \cdot 1 \end{bmatrix} \in \mathcal{R}^{n \times n}$$

and (with the use of (10)*):*

$$\mathbf{G}_n = \frac{H^2}{4} \cdot \begin{bmatrix} \frac{9}{7} & \frac{6}{20} & & & & \\ \frac{3}{7} & \frac{9}{20} & \frac{6}{80} & & & \\ & \frac{3}{20} & \frac{9}{80} & \ddots & & \\ & & \frac{3}{80} & \ddots & \frac{6}{20 \cdot 4^{n-3}} & \\ & & & \ddots & \frac{9}{20 \cdot 4^{n-3}} & \frac{6}{56 \cdot 2^n} \\ & & & & \frac{3}{20 \cdot 4^{n-3}} & \frac{9}{56 \cdot 2^n} \end{bmatrix} \in \mathcal{R}^{n \times n}.$$

At its turn this leads to:

$$\mathbf{A}_n\mathbf{G}_n = \begin{bmatrix} 1 & 0 & -\frac{1}{20} & & & \\ 0 & 1 & 0 & \ddots & & \\ -\frac{48}{21} & 0 & \ddots & & -\frac{1}{20} & \\ & -\frac{32}{10} & & \ddots & 0 & -\frac{1}{28} \\ & & \ddots & & 1 & 0 \\ & & & -\frac{32}{10} & 0 & 1 \end{bmatrix}, \tag{13}$$

and

$$\mathbf{G}_n\mathbf{A}_n = \begin{bmatrix} \frac{31}{35} & \frac{12}{35} & -\frac{4}{5} & & & \\ -\frac{6}{35} & \frac{39}{35} & 0 & \ddots & & \\ -\frac{1}{5} & 0 & 1 & & -\frac{4}{5} & \\ & -\frac{1}{5} & & 1 & 0 & -\frac{4}{5} \\ & & \ddots & 0 & \frac{39}{35} & -\frac{12}{35} \\ & & & -\frac{1}{5} & \frac{6}{35} & \frac{31}{35} \end{bmatrix}. \tag{14}$$

These two products are almost Toeplitz and are approximated with the Toeplitz matrices:

$$\mathbf{B}_n^{\mathrm{r}} = \begin{bmatrix} 1 & 0 & -\frac{1}{20} & & & \\ 0 & 1 & 0 & \ddots & & \\ -\frac{32}{10} & 0 & \ddots & & -\frac{1}{20} & \\ & -\frac{32}{10} & & \ddots & \ddots & -\frac{1}{20} \\ & & \ddots & & 1 & 0 \\ & & & -\frac{32}{10} & 0 & 1 \end{bmatrix}$$

and

$$\mathbf{B}_n^{\mathrm{l}} = \begin{bmatrix} 1 & 0 & -\frac{4}{5} & & & \\ 0 & 1 & 0 & \ddots & & \\ -\frac{1}{5} & 0 & \ddots & & -\frac{4}{5} & \\ & -\frac{1}{5} & & 1 & \ddots & -\frac{4}{5} \\ & & \ddots & & 1 & 0 \\ & & & -\frac{1}{5} & 0 & 1 \end{bmatrix}. \tag{15}$$

In this case

$$\mathbf{B}_{n/2}^{\mathrm{l}} = \begin{bmatrix} 1 & -\frac{4}{5} & & \\ -\frac{1}{5} & 1 & \ddots & \\ & \ddots & \ddots & -\frac{4}{5} \\ & & -\frac{1}{5} & 1 \end{bmatrix}.$$

which via 9 implies that

$$\sigma(\mathbf{B}_n^{\mathrm{l}}) = \sigma(\mathbf{B}_{n/2}^{\mathrm{l}}) = \left(1 + \frac{4}{5}\cos(\frac{i\pi}{n+1})\right)_{i=1}^{n/2} \subset (\frac{1}{5}, \frac{9}{5}), n \to \infty.$$

4 The condition number of the preconditioned systems

The calculation of $\kappa_2(\mathbf{B}_n^{\mathrm{r}})$ and $\kappa_2(\mathbf{B}_n^{\mathrm{l}})$ is not trivial because these matrices are not symmetric. For the latter matrix we resort to the calculation of an inverse presented in [10, Corr. 5.5]:

Theorem 4. *Let $\Omega = (0,1)$ be partitioned as $0 = \tilde{x}_0 < \tilde{x}_1 < \ldots < \tilde{x}_{n+1} = 1$. Let $\tilde{h}_i = \tilde{x}_{i+1} - \tilde{x}_i$. Let $\tilde{\mathbf{A}}_n \in \mathcal{R}^{n\times n}$ be the matrix*

$$\tilde{\mathbf{A}}_n = [-\frac{2\tilde{h}_i}{\tilde{h}_i + \tilde{h}_{i-1}}, 2, -\frac{2\tilde{h}_{i-1}}{\tilde{h}_i + \tilde{h}_{i-1}}]. \tag{16}$$

Then (see [10, Corr. 5.5]):

$$[\tilde{\mathbf{A}}_n^{-1}]_{i,j} = \begin{array}{lll} \dfrac{\tilde{h}_j + \tilde{h}_{j-1}}{2\tilde{h}_j\tilde{h}_{j-1}} & (1-\tilde{x}_j)\tilde{x}_i, & i \leq j \\ \dfrac{\tilde{h}_j + \tilde{h}_{j-1}}{2\tilde{h}_j\tilde{h}_{j-1}} & \tilde{x}_j(1-\tilde{x}_i), & i \geq j. \end{array} \tag{17}$$

The trick is to establish that $\mathbf{B}_n^{\mathrm{l}}$ is of the form $\tilde{\mathbf{A}}_n$, i.e., is related to the discretization of the Laplace operator on a different grid (just the odd or the even points, with grow-factor ψ^2). This is done as follows.

Lemma 1. *Let $\psi \leq 1$ and $\Psi := \psi^2$. Recall that due to* (12) *there exist b_{-1} and b_1 such that*

$$\mathbf{B}_{n/2}^{\mathrm{l}} = [\frac{b_{-1}}{2}, 1, \frac{b_1}{2}].$$

Then $\mathbf{B}_{n/2}^{\mathrm{l}}$ is of the form (16)*, and*

$$[(\mathbf{B}_{n/2}^{\mathrm{l}})^{-1}]_{i,j} = \begin{array}{lll} \Psi^{-j}(1+\Psi)(1+\Psi+\ldots+\Psi^{n/2}) & (1-\tilde{x}_j)\tilde{x}_i, & i \leq j \\ \Psi^{-j}(1+\Psi)(1+\Psi+\ldots+\Psi^{n/2}) & \tilde{x}_j(1-\tilde{x}_i), & i \geq j. \end{array}$$

Proof. First we show that $\mathbf{B}_{n/2}^{\mathrm{l}}$ is of a similar form as $\tilde{\mathbf{A}}_n$. In particular we should have

$$b_{-1} = -\frac{2\tilde{h}_{i+1}}{\tilde{h}_{i+1} + \tilde{h}_i} \text{ and } b_1 = -\frac{2\tilde{h}_i}{\tilde{h}_{i+1} + \tilde{h}_i}, \tag{18}$$

cf. (16). The required form in (18) holds if

$$\frac{b_{-1}}{b_1} = \frac{\tilde{h}_{i+1}}{\tilde{h}_i}.$$

In our case

$$b_{-1} = -\frac{2\psi^2}{1+\psi^2}, \qquad b_1 = -\frac{2}{1+\psi^2} \Rightarrow \frac{\tilde{h}_{i+1}}{\tilde{h}_i} = \frac{-\frac{2\psi^2}{1+\psi^2}}{-\frac{2}{1+\psi^2}} = \psi^2.$$

Let $\Psi := \psi^2$ and $m := n/2$. Then

$$1 = \sum_{i=1}^{m+1} \tilde{h}_i = \tilde{h}_1 \sum_{i=0}^{m} \Psi^i = \tilde{h}_1 \frac{1-\Psi^{m+1}}{1-\Psi} \implies \tilde{h}_i = \frac{1-\Psi}{1-\Psi^{m+1}}\Psi^{i-1}$$

and related

$$\tilde{x}_i = \frac{1-\Psi}{1-\Psi^{m+1}} \sum_{j=1}^{i} \Psi^{j-1} = \frac{1-\Psi}{1-\Psi^{m+1}} \frac{1-\Psi^i}{1-\Psi} = \frac{1-\Psi^i}{1-\Psi^{m+1}}.$$

for all $i = 0, \ldots, n$. □

Having acquired expressions for h_i and $\tilde{x}_i$, we use (17) to compute $(\mathbf{B}^{\mathrm{l}}_{n/2})^{-1}$:

Lemma 2.

$$\|(\mathbf{B}^{\mathrm{l}}_n)^{-1}\|_\infty \leq \frac{1+\Psi}{1-\Psi}(\frac{n}{2}+1), \quad \kappa_2(\mathbf{B}^{\mathrm{l}}_n) = O(n), \;\; n\to\infty. \tag{19}$$

Proof. Let $m := n/2$. The i-th row sum of $\|(\mathbf{B}^{\mathrm{l}}_{n/2})^{-1}\|_\infty$ is:

$$r_i = \frac{1+\Psi}{1-\Psi-\Psi^{m+1}+\Psi^{m+2}}((m+1)(1-\Psi^i)+i(\Psi^{m+1}-1)).$$

One finds, for the first factor (which is independent of i):

$$\lim_{n\to\infty} \frac{1+\Psi}{1-\Psi-\Psi^{m+1}+\Psi^{m+2}} = \frac{1+\Psi}{1-\Psi}$$

and for the second factor:

$$\begin{aligned}(m+1)(1-\Psi^i)+i(\Psi^{m+1}-1) &\leq (m+1)(1-\Psi^i)\\ &\leq m+1.\end{aligned}$$

This leads to

$$\|(\mathbf{B}^{\mathrm{l}}_n)^{-1}\|_\infty = \|(\mathbf{B}^{\mathrm{l}}_{n/2})^{-1}\|_\infty \leq \frac{1+\Psi}{1-\Psi}(\frac{n}{2}+1).$$

The second claim follows from $\|\mathbf{B}^{\mathrm{l}}_n\|_\infty = 2$ in combination with the first claim. □

The matrix $\mathbf{B}^{\mathrm{r}}_n$ does not satisfy the conditions from [10, Corr. 5.5] because its rows sums $i = 2,\ldots,n-1$ are not zero. However, an extra scaling with a diagonal matrix will solve this problem. Nevertheless, our bound is not optimal, as numerical results in Table 1 show.

Lemma 3.

$$\kappa_2(\mathbf{B}^{\mathrm{r}}_{n/2}) \doteq O(\frac{n}{2}\left(\frac{1}{\psi^2}\right)^n), \;\; n\to\infty. \tag{20}$$

Proof. Observe that $\mathbf{B}^{\mathrm{r}}_n$ can be permuted to yield

$$\bar{\mathbf{B}}^{\mathrm{r}}_n = \begin{bmatrix} \mathbf{B}^{\mathrm{r}}_{n/2} & \\ & \mathbf{B}^{\mathrm{r}}_{n/2}\end{bmatrix}$$

where

$$\mathbf{B}^{\mathrm{r}}_{n/2} = [-\frac{1}{\psi^2(1+\psi^2)}, 1, -\frac{\psi^4}{1+\psi^2}].$$

Let $\mathbf{D}_{n/2}$ be a diagonal matrix defined by

$$[\mathbf{D}_{n/2}]_{i,i} := \left(\frac{1}{\psi^2}\right)^{i-1}, \qquad i = 1,\ldots,n/2. \tag{21}$$

Then

$$\mathbf{D}_{n/2}^{-1}\mathbf{B}_{n/2}^{\mathrm{r}}\mathbf{D}_{n/2} = (\mathbf{B}_{n/2}^{\mathrm{l}})^{T},$$

which implies

$$\mathbf{B}_{n/2}^{\mathrm{r}} = \mathbf{D}_{n/2}(\mathbf{B}_{n/2}^{\mathrm{l}})^{T}\mathbf{D}_{n/2}^{-1}. \tag{22}$$

Since $\kappa_2(\mathbf{B}_{n/2}^{\mathrm{l}}) = O(n/2)$ and $\kappa_2(\mathbf{D}_{n/2}) = \psi^{-n}$ one obtains the desired result. □

5 Numerical results

The numerical results are provided for a range of different tensor refined grids on the unit d-cube domain $\Omega \subset \mathcal{R}^d$ for the convection diffusion equation with Dirichlet boundary conditions (1). For all tests we use tensor-grids defined with the use of (3), $n_1 = n_2 = n_3$ and $\psi = 1/2$. For $d = 2$ dimensions this leads to the grids T-1, T-2 and T-n in Figures 1 and 2. In three dimensions the grid T-3 is refined into the corner $(1, 1, 1)$.

For all numerical tests, we chose a convection vector $\mathbf{v} = c \cdot \mathbf{b} \in \mathcal{R}^d$ which follows the refined grid: If there is grid refinement towards to the $+1$ corner in the x_i direction then $b_i = 1$ and $b_i = 0$ elsewise. If there is refinement in all spatial directions we use $b_i = 1$ for all directions – for comparison purposes. The scale factor $c = 0$ (no convection) or $c = 100$ (convection).

For our tests the iteration count did not depend on the smoothness of the right hand side of the linear system. To determine this we used solutions such as $x \longrightarrow \frac{\pi}{2} + \arctan(100(x-1))$ through $x, y, z \mapsto 16 \cdot x(1-x)y(1-y)z(1-z) + x^2 + y^2 + z^2$. Though GMRES [14] is applicable CGS [15] takes less iterations and is our iterative method of choice.

All tables contain results for a range of different preconditioners constructed with the approximate inverse and ILU(0) factorization. CGS-k stands for CGS applied to:

CGS-1: The original system of equations:

$$\mathbf{A}_n\mathbf{x}_n = \mathbf{b}_n;$$

CGS-2: The right preconditioned system:

$$\mathbf{A}_n\mathbf{G}_n\mathbf{y}_n = \mathbf{b}_n; \;\; \mathbf{x}_n = \mathbf{G}_n\mathbf{y}_n,$$

CGS-3: The left preconditioned system:

$$\mathbf{G}_n\mathbf{A}_n\mathbf{x}_n = \mathbf{G}_n\mathbf{b}_n;$$

CGS-4: The original system of equations with $\mathbf{K}_1$ ILU(0) factorization of $\mathbf{A}_n$:

$$\mathbf{K}_1^{-1}\mathbf{A}_n\mathbf{x}_n = \mathbf{K}_1^{-1}\mathbf{b}_n;$$

CGS-5: The right preconditioned system with $\mathbf{K}_2$ ILU(0) factorization of $\mathbf{A}_n\mathbf{G}_n$:

$$\mathbf{K}_2^{-1}\mathbf{A}_n\mathbf{G}_n\mathbf{y}_n = \mathbf{K}_2^{-1}\mathbf{b}_n, \;\; \mathbf{x}_n = \mathbf{G}_n\mathbf{y}_n;$$

CGS-6: The left preconditioned system with $\mathbf{K}_3$ ILU(0) factorization of $\mathbf{G}_n\mathbf{A}_n$:

$$\mathbf{K}_3^{-1}\mathbf{G}_n\mathbf{A}_n\mathbf{x}_n = \mathbf{K}_3^{-1}\mathbf{G}_n\mathbf{b}_n.$$

In all cases, CGS is stopped when the absolute residual is smaller than 10^{-12} – initial residuals are between $1/2$ and 2. A bar ("-") denotes that no convergence is obtained, probably due to the unscaled standard ILU(0) preconditioner (see [13] for scaled ILUs). The standard incomplete LU(0) factorization (see for instance [1]) depends on the numbering of the degrees of freedom which is in all cases left to right and bottom to top.

For the case $d = 1$, $c = 0$ and $\psi = 1/2$ one finds Toeplitz approximations $\mathbf{B}_n^{\mathrm{r}}$, $\mathbf{B}_n^{\mathrm{l}}$ and $\mathbf{D}_n$ with the use of (11), (12) and (21):

$$\mathbf{B}_n^{\mathrm{r}} = [-\frac{32}{10},\; 1,\; -\frac{1}{20}], \quad \mathbf{B}_n^{\mathrm{l}} = [-\frac{1}{5},\; 1,\; -\frac{4}{5}], \quad [\mathbf{D}_n]_{i,i} = 4^{i-1}.$$

The numerical results for $\kappa_2(\mathbf{B}_n^{\mathrm{l}})$ in Table 1 corroborate our theoretical result (19). They also indicate that the theoretical result (20) is not sharp and suggest instead the better bound

$$\kappa_2(\mathbf{B}_{n/2}^{\mathrm{r}}) \sim O\left(\frac{1}{\psi}\right)^n), \;\; n \to \infty. \tag{23}$$

For dimensions $d = 2$ and $d = 3$ the results in Tables 2, 3, 4 and 5 indicate that the real approximate inverse systems $\mathbf{A}_n\mathbf{G}_n$ and $\mathbf{G}_n\mathbf{A}_n$ behave just as the approximations $\mathbf{B}_n^{\mathrm{r}}$ and $\mathbf{B}_n^{\mathrm{l}}$ – *even in the convection dominant case.* They also show that an additional ILU(0) factorization applied to $\mathbf{A}_n\mathbf{G}_n$ and $\mathbf{G}_n\mathbf{A}_n$ can lead to an (almost) optimally conditioned system.

The numerical results for other combinations of refinement such as in Figure 2 are all similar, see [7].

References

[1] O. Axelsson and V.A. Barker. *Finite Element Solution of Boundary Value Problems*. Academic Press, Orlando, Florida, 1984.

[2] O. Axelsson and P.S. Vassilevski. Algebraic multilevel preconditioning methods i. *Numerische Mathematik*, 56:157–177, 1989.

[3] O. Axelsson and P.S. Vassilevski. Algebraic multilevel preconditioning methods ii. *SIAM Journal on Numerical Analysis*, 27:1569–1590, 1990.

[4] M. Benzi and M. Tuma. A sparse approximate inverse preconditioner for nonsymmetric linear systems. *SIAM Journal on Scientific Computing*, 19:968–994, 1998.

[5] A. Böttcher and B. Silbermann. *Introduction to large truncated Toeplitz matrices*. Springer, New York, 1999.

[6] J. D. F. Cosgrove, J. C. Diaz, and A. Griewank. Approximate inverse preconditionings for sparse linear systems. *International Journal of Computer Mathematics*, 44:91–110, 1992.

[7] W. D. Drenth. *A platform for numerical computations with special application to preconditioning*. PhD thesis, Eindhoven University of Technology.

[8] L. Yu. Kolotilina and A. Yu. Yeremin. Factorized sparse approximate inverse preconditioning i: Theory. *SIAM Journal onf Matrix Analysis and Applications*, 14:45–58, 1993.

[9] S.D. Margenov and J.M. Maubach. Optimal algebraic multilevel preconditioning for local refinement along a line. *Journal of Numerical Linear Algebra with Applications*, 2:347–362, 1995.

[10] R. M. M. Mattheij and M. D. Smooke. Estimates for the inverse of tridiagonal matrices arising in boundary-value problems. *Linear Algebra and its Applications*, 73:33–57, 1986.

[11] J. Maubach. Parallel local bisection refinement for n-dimensional simplicial grids. In *ISBN 1-880843-02-1 Proceedings of the ISMM International Conference on Parallel and Distributed Computing and Systems, (ed. R. Melhem), Pittsburgh, Pennsylvania, U.S.A.*, pages 310–312, 1992.

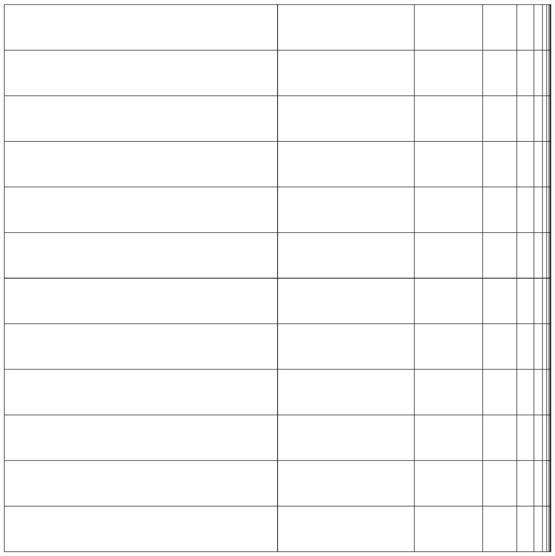
[12] J. Maubach. Local bisection refinement for n-simplicial grids generated by reflections. *SIAM Journal on Scientific Computing*, 16:210–227, 1995.

[13] Y. Saad. *Iterative methods for sparse linear systems*. PWS, 1996.

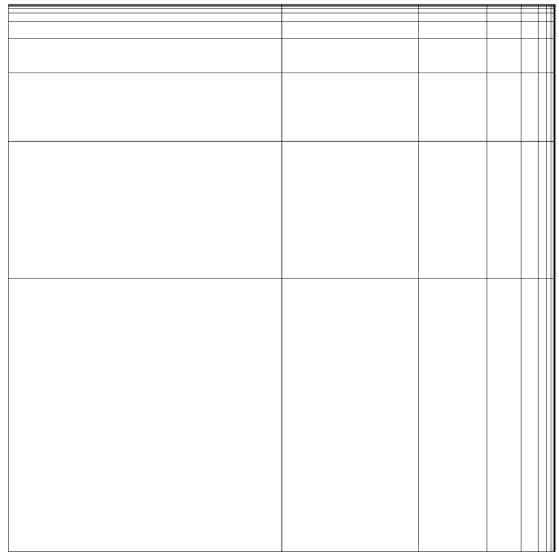
[14] Y. Saad and M. H. Schulz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7:856–869, 1986.

[15] P. Sonneveld. CGS, a fast Lanczos-type solver for nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 10:36–52, 1989.

[16] W.-P. Tang. Toward an effective sparse approximate inverse preconditioner. *SIAM Journal on Matrix Analysis and Application*, 20:970–986, 1999.

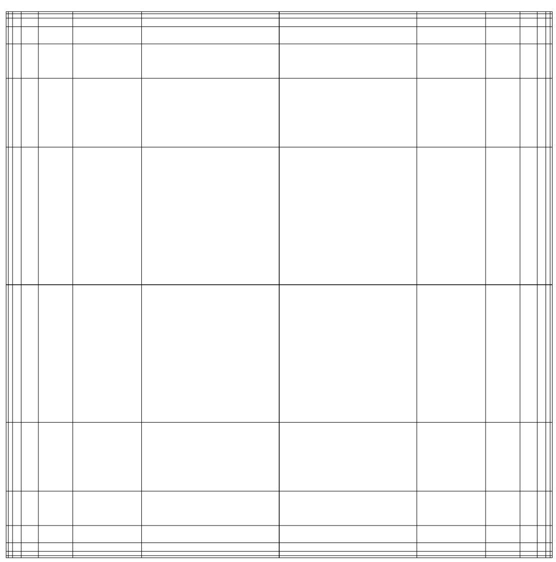


(a) Tensor grid T-1 with refinement to $x = 1$, $\psi = 1/2$.



(b) Tensor grid T-2 with refinement to $x = 1$ and $y = 1$, $\psi = 1/2$.

Figure 1: Tensor grids with different refinements.



(a) Tensor grid T-n with refinement into all corners, $\psi = 1/2$.

Figure 2: Tensor grid T-n with refinement into all corners, $\psi = 1/2$.

n	5	6	7	8	9	10	11
$\kappa_2(\mathbf{B}_n^{\rm r})$	1086.09	4390.73	$1.77{\cdot}10^4$	$7.10{\cdot}10^4$	$2.85{\cdot}10^5$	$1.14{\cdot}10^6$	$4.58{\cdot}10^6$
$\kappa_2(\mathbf{B}_n^{\rm l})$	9.20	11.38	13.55	15.71	17.87	20.03	22.18
$\kappa_2(\mathbf{D}_n)$	4^4	4^5	4^6	4^7	4^8	4^9	4^{10}
$\kappa_2(\mathbf{B}_n^{\rm r})/\kappa_2(\mathbf{D}_n)$	4.24	4.29	4.32	4.34	4.35	4.36	4.37

Table 1: Results for the grid T-1 (1-d) without convection $c = 0$.

n (n_k)	16 (4)	25 (5)	36 (6)	49 (7)	64 (8)	81 (9)	100 (10)	121 (11)
$\kappa_2(\mathbf{A}_n)$	121.61	506.14	2066.41	8352.85	$3.36{\cdot}10^4$	$1.35{\cdot}10^5$	$5.40{\cdot}10^5$	$2.16{\cdot}10^6$
$\kappa_2(\mathbf{A}_n\mathbf{G}_n)$	8.03	20.89	40.12	92.64	172.70	391.37	718.89	1612.62
$\kappa_2(\mathbf{G}_n\mathbf{A}_n)$	2.78	3.83	4.80	5.90	6.92	8.03	9.08	10.19
CGS-1	11	21	37	-	-	-	-	-
CGS-2	9	13	14	16	17	19	20	23
CGS-3	8	11	11	13	15	16	17	18
CGS-4	-	-	-	-	-	-	-	-
CGS-5	5	7	-	-	-	-	-	-
CGS-6	4	5	5	5	5	6	6	6

Table 2: Results for the grid T-2 without convection $c = 0$.

n (n_k)	16 (4)	25 (5)	36 (6)	49 (7)	64 (8)	81 (9)	100 (10)	121 (11)
$\kappa_2(\mathbf{A}_n)$	41.81	129.53	432.80	1553.95	5866.31	$2.28{\cdot}10^4$	$8.99{\cdot}10^4$	$3.57{\cdot}10^5$
$\kappa_2(\mathbf{A}_n\mathbf{G}_n)$	23.23	54.07	118.92	255.26	523.16	1072.20	2152.34	4370.88
$\kappa_2(\mathbf{G}_n\mathbf{A}_n)$	3.78	4.67	5.42	6.05	6.56	6.97	7.33	7.70
CGS-1	11	17	30	-	-	-	-	-
CGS-2	6	8	10	12	14	17	19	20
CGS-3	6	8	9	10	12	13	16	16
CGS-4	-	-	-	-	-	-	-	-
CGS-5	-	-	-	-	-	-	-	-
CGS-6	3	3	4	5	5	5	5	6

Table 3: Results for the grid T-2 with convection $c = 100$.

n (n_k)	27 (3)	125 (5)	343 (7)	729 (9)	1331 (11)	2197 (13)	3375 (15)	4913 (17)
$\kappa_2(\mathbf{A}_n)$	28.36	512.47	8474.53	$1.37{\cdot}10^5$	$2.19{\cdot}10^6$	$3.51{\cdot}10^7$	$5.62{\cdot}19^8$	$8.99{\cdot}10^9$
$\kappa_2(\mathbf{A}_n\mathbf{G}_n)$	3.11	15.16	68.23	289.31	1197.38	4881.77	$1.98{\cdot}10^4$	$7.95{\cdot}10^4$
$\kappa_2(\mathbf{G}_n\mathbf{A}_n)$	1.85	3.63	5.75	7.92	10.09	12.25	14.41	16.57
CGS-1	10	-	-	-	-	-	-	-
CGS-2	8	13	17	22	22	25	30	34
CGS-3	7	10	13	16	18	18	19	20
CGS-4	7	-	-	-	-	-	-	-
CGS-5	4	8	-	-	-	-	-	-
CGS-6	4	5	6	6	6	6	6	6

Table 4: Results for the grid T-3 without convection $c = 0$.

n (n_k)	27 (3)	125 (5)	343 (7)	729 (9)	1331 (11)	2197 (13)	3375 (15)	4913 (17)
$\kappa_2(\mathbf{A}_n)$	16.07	166.60	2143.91	$3.23 \cdot 10^4$	$5.11 \cdot 10^5$	$8.16 \cdot 10^6$	$1.31 \cdot 10^8$	$2.09 \cdot 10^9$
$\kappa_2(\mathbf{A}_n\mathbf{G}_n)$	8.16	51.83	271.05	1143.05	4627.43	$1.87 \cdot 10^4$	$7.53 \cdot 10^4$	$3.02 \cdot 10^5$
$\kappa_2(\mathbf{G}_n\mathbf{A}_n)$	3.26	6.35	8.77	10.18	10.93	11.59	12.49	13.57
CGS-1	10	36	-	-	-	-	-	-
CGS-2	5	9	13	20	23	26	31	35
CGS-3	5	8	13	16	18	21	22	23
CGS-4	4	-	-	-	-	-	-	-
CGS-5	3	6	-	-	-	-	-	-
CGS-6	3	4	5	6	6	7	7	7

Table 5: Results for the grid T-3 with convection $c = 100$.