Cholesky factorisations of linear systems coming from a finite difference method applied to singularly perturbed problems

Thai Anh Nhan and Niall Madden
The Boundary and Interior Layers - Computational & Asymptotic Methods

Charles University in Prague, September 2014
Outline...

1. The model problem

2. Cholesky factorization on a uniform mesh
   - Analysis
   - An application: Incomplete Cholesky with threshold

3. Cholesky factorization on a boundary layer-adapted mesh

4. Conclusion
Problem:
We consider the singularly perturbed two dimensional reaction-diffusion problem:
\[-\varepsilon^2 \Delta u + b(x, y)u = f(x, y), \text{ on } \Omega = (0, 1)^2, \text{ and } u(\partial \Omega) = g(x, y),\] (1)
where \(\varepsilon\) is a “perturbation parameter”, \(0 < \varepsilon \ll 1\), \(g(x, y)\), \(b(x, y)\) and \(f(x, y)\) are some given functions such that \(b(x, y) \geq \beta^2 > 0\). It is known that a standard finite difference method applied to (1) on a boundary layer-adapted mesh with \(N\) intervals in each direction yields a parameter robust approximation (cf. Clavero et al. (2005) and Linß (2010)). This method leads to a linear system of equations which must be solved. We write the system as
\[AU = F,\] (2)
where \(A\) is a banded, symmetric and positive definite \((N - 1)^2 \times (N - 1)^2\) matrix.
Motivation
We wish to investigate the Cholesky factorization of matrices in which the magnitudes of diagonal entries are dominant compared to the off-diagonal entries such as the coefficient matrix $A$ defined in (2). It is observed by MacLachlan and Madden (2013) that

- The performance of direct solvers (Cholesky factorization) for sparse matrices of this type depends not only on the location of nonzero of the matrix, but also their values.
- Factorization process results in many extremely small numbers, so-called subnormal and underflow numbers, in the Cholesky factors. In practice, this affects the computational speed considerably, and so the amount of time required to solve these linear systems depends badly on the perturbation parameter.
- The fill-in entries in the Cholesky factorisation decay exponentially away from the main diagonal. Thus, for small $\varepsilon$ and large $N$, the Cholesky factors contain many subnormal and underflow numbers. We want to provide a mathematical justification for this phenomenon, and also study how it can be exploited to our advantages.
Cholesky factorization on a uniform mesh

Initially, we consider the standard central finite difference discretization of the model problem (1) on uniform mesh with \(N\) intervals on each direction. The equally spaced stepsize is denoted by \(h = N^{-1}\) and we will suppose that \(\varepsilon \ll h\), which is reasonable in practice. Then the discrete matrix \(A\) can be written as the following 5-point stencil

\[
A = \begin{pmatrix}
-\varepsilon^2 & 4\varepsilon^2 + h^2 b_{i,j} & -\varepsilon^2 \\
-\varepsilon^2 & 0 & -\varepsilon^2 \\
-\varepsilon^2 & -\varepsilon^2 & -\varepsilon^2
\end{pmatrix} = \begin{pmatrix}
-\varepsilon^2 & 0 & -\varepsilon^2 \\
0 & O(h^2) & 0 \\
-\varepsilon^2 & -\varepsilon^2 & -\varepsilon^2
\end{pmatrix},
\]

since \((4\varepsilon^2 + h^2 b_{i,j}) = O(h^2)\), where we write \(f(\cdot) = O(g(\cdot))\) if there exists positive constants \(C_0, C_1\) independent of variables such that \(C_0|g(\cdot)| \leq f(\cdot) \leq C_1|g(\cdot)|\). This problem is related to the region in the interior of a boundary layer-adapted mesh for problems (1) where the local mesh width of order \(O(N^{-1})\). Therefore, the Cholesky factorization in the interior region is essential the same as that of the discretized matrix \(A\) on a uniform mesh.
There are several ways to implement Cholesky factorization of a symmetric positive definite matrix. Below, we present in Algorithm 1 a version adapted from Golub and Van Loan, *Matrix Computations* (1996). It computes a lower triangular matrix $L$ such that $A = LL^T$. We will follow Matlab notation by denoting $A = [a(i,j)]$ and $L = [l(i,j)]$.

**Algorithm 1: Cholesky factorization**

for $j = 1 : n$
  if $j = 1$
    for $i = j : n$
      $l(i, j) = \frac{a(i, j)}{\sqrt{a(j, j)}}$
    end
  elseif $(j > 1)$
    for $i = j : n$
      $l(i, j) = \frac{a(i, j) - \sum_{k=1}^{j-1} l(i, k)l(j, k)}{\sqrt{a(j, j)}}$ (line 8)
    end
  end
end
Figures below show the structures of the coefficient matrix $A$ (on the left) and Cholesky factor $L$ (on the right) when $N = 8$.

**Figure:** The matrix $A$ (left) and Cholesky factor $L$ (right).
Figures below show the structures of the coefficient matrix $A$ (on the left) and Cholesky factor $L$ (on the right) when $N = 8$.

![Figure: The matrix $A$ (left) and Cholesky factor $L$ (right).](image)

The fill-in entries are colored in red.

**Question:** can we estimate the magnitudes of these fill-in entries in terms of $\varepsilon$ and $h$?
**Terminology** (based on Chapter 10, Y. Saad (2003))

To analyse the magnitudes of the fill-in entries, we will denote

- The distinct sets $L^{[0]}$, $L^{[1]}$, $\ldots$, $L^{[m]}$ where all entries of the same magnitude belong to the same set in a sense that if $l^{[k]}$ is the magnitude of entry in $L^{[k]}$, then $l(i, j) \in L^{[k]}$ if and only if $l(i, j) \in O(l^{[k]})$. We shall see that these sets are quite distinct, meaning that $l^{[k]} \gg l^{[k+1]}$.

- All the nonzero $a(i, j)$ of the original matrix $A$ belong to $L^{[0]}$.

- The Cholesky process begins with zero matrix $L$. We define these initial zeros belong to $L^{[\infty]}$.

Suppose that $l(i, j) \in L^{[p_{i,j}]}$ (with $p_{i,j} = \infty$ from the start). At each sweep through the algorithm, new values of $l(i, j)$ are computed, and so $p_{i,j}$ is modified. Then, from line 8 in Algorithm 1, we can see that the $p_{i,j}$ is updated by

$$p_{i,j} = \begin{cases} 0, & \text{if } a(i, j) \neq 0, \\ \min\{p_{i,1} + p_{j,1} + 1, p_{i,2} + p_{j,2} + 1, \ldots, p_{i,j-1} + p_{j,j-1} + 1\}, & \text{otherwise}. \end{cases}$$  \hspace{1cm} (4)
Let us denote \( m = N - 1 \), the band width of \( A \) defined in (3). Then our analysis has shown that, the fill-in entries from Cholesky factorization have the following structure. The first \( 2m \) rows of matrix \( L \) is shown below in which, for simplicity, the entries belonging to \( L^{[k]} \) are denoted by \([k]\) and the entries that corresponding to nonzero entries of original matrix are written in terms of their magnitude.

\[
\begin{pmatrix}
O(h) & O(h) & O(h) \\
O(\varepsilon^2/h) & O(h) & O(h) \\
O(\varepsilon^2/h) & O(\varepsilon^2/h) & O(h) \\
\vdots & \vdots & \vdots \\
O(\varepsilon^2/h) & [1] & \ldots & [m-2] & O(h) \\
O(\varepsilon^2/h) & [1] & \ldots & [m-2] & O(h) \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
O(\varepsilon^2/h) & [1] & \ldots & [3] & O(h) \\
O(\varepsilon^2/h) & [m] & \ldots & [3] & O(h) \\
\end{pmatrix}
\]

**Figure:** The submatrix \( L(1:2m, 1:2m) \).

It turns out that the remaining fill-in entries in \( L \) have the same structure as the fill-in entries in submatrix \( L(m+1:2m, 1:2m) \).
We present a closer look on the submatrix $L(m + 1 : 2m, 1 : m)$.

$$
\begin{pmatrix}
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\mathcal{O}(\varepsilon^2/h) & [1] & \mathcal{O}(\varepsilon^2/h) & [1] & \mathcal{O}(\varepsilon^2/h) & \mathcal{O}(\varepsilon^2/h)
\end{pmatrix}
$$

**Figure:** The submatrix $L(m + 1 : 2m, 1 : m)$. 
And the submatrix $L(m + 1 : 2m, m + 1 : 2m)$.

$$
\begin{pmatrix}
\mathcal{O}(h) & \mathcal{O}(h) & \mathcal{O}(h) \\
\mathcal{O}(\varepsilon^2/h) & \mathcal{O}(h) & \mathcal{O}(h) \\
[3] & \mathcal{O}(\varepsilon^2/h) & \mathcal{O}(h) \\
[4] & [3] & \mathcal{O}(\varepsilon^2/h) & \mathcal{O}(h) \\
\vdots & \vdots & \ddots & \ddots & \ddots \\
[m - 1] & [m - 2] & \ldots & [3] & \mathcal{O}(\varepsilon^2/h) & \mathcal{O}(h) \\
\end{pmatrix}
$$

**Figure:** The partition $L(m + 1 : 2m, m + 1 : 2m)$. 

Thái Anh Nhan (NUI Galway, Ireland)
We have shown that

\[ \text{Theorem 1} \]

Let \( \delta = \varepsilon / h \). If \( l(i, j) \in L^{[k]} \), then the magnitude \( l^{[k]} \) of \( l(i, j) \) is given as follows,

\[ l^{[k]} = O \left( \delta^{2(k+1)} h \right). \quad (5) \]

The formulation given in (5) tells us why the values of fill-in entries decay exponentially with respect to \( k \). In practice, for a reaction-dominated problem, we usually have \( \varepsilon \ll h \). Hence, when \( \varepsilon \) decreases and the mesh parameter \( N \) increases, the fill-in entries tend to zero rapidly. This fact also suggests that an Incomplete Cholesky factorization would be an excellent preconditioner for iterative solvers (not discussed here).
To demonstrate Theorem 1, we plot the magnitude of the entries of the vector $L(m + 1, 1 : m)$ with various values of $\varepsilon$ and $N = 128$.

It clearly shows that the magnitude of the fill-in entries decay exponentially, which is in agreement with our result.

Furthermore, when $\varepsilon = 10^{-6}$, there is only first 41 entries of the vector $L(128, 1 : 127)$ can be plotted because the magnitudes of the other entries are less than the smallest number in IEEE standard double precision, and so they are flushed to zero.

**Figure:** Semilog plot of the magnitude of $L(128, 1 : 127)$ with various $\varepsilon$. 
Furthermore, we can give the exact number of the fill-in entries associated with their magnitudes.

**Theorem 2**

Let \(|L^{[k]}|\) denote the number of elements in the set \(L^{[k]}\). Then

\[
\begin{align*}
\sum_{k=1}^{m} |L^{[k]}| &\quad \text{sum of the number of fill-in entries} \\
\sum_{k=2}^{m} |L^{[k]}| &\quad (m-2)(m-1)^2 \\
\sum_{k=p}^{m} |L^{[k]}|, \forall p \geq 3 &\quad (m-1)(m-p+1)^2
\end{align*}
\]

Table: The sum of the number of fill-in entries associated with their magnitudes.
**Numerical experiments**

To validate Theorem 2, we compare the sum of all fill-in entries which belong to $L[k], k \geq p$ given by Theorem 2 and the actual number of fill-in entries that are less than the magnitude $l[p]$ is computed by Matlab.

<table>
<thead>
<tr>
<th>p</th>
<th>$\delta^{2(p+1)}h$</th>
<th>Actual number</th>
<th>Theorem 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.62e-07</td>
<td>238328</td>
<td>238328</td>
</tr>
<tr>
<td>2</td>
<td>1.07e-09</td>
<td>234484</td>
<td>234484</td>
</tr>
<tr>
<td>3</td>
<td>4.39e-12</td>
<td>230702</td>
<td>230702</td>
</tr>
<tr>
<td>4</td>
<td>1.80e-14</td>
<td>223200</td>
<td>223200</td>
</tr>
<tr>
<td>5</td>
<td>7.37e-17</td>
<td>215822</td>
<td>215822</td>
</tr>
<tr>
<td>6</td>
<td>3.02e-19</td>
<td>208568</td>
<td>208568</td>
</tr>
</tbody>
</table>

**Table:** The actual and estimated numbers of fill-in entries associated with their magnitudes when $\varepsilon = 10^{-3}$ and $N = 2^6$. 
An application: Incomplete Cholesky with threshold

- Note that the matrix $A$ is symmetric and positive definite, so there exists a complete Cholesky factorization:

$$A = LL^T$$

where $L$ is a lower triangular matrix.

- **Incomplete Cholesky**: set some entries of $L$ to zeros (or, rather, don’t compute them at all), and so $A \approx \tilde{L}\tilde{L}^T$ where $\tilde{L}$ is the Incomplete Cholesky factor.

- $\tilde{L}$ is quite inexpensive to compute compared to $L$. Since $A \approx \tilde{L}\tilde{L}^T$, it can be used as a preconditioner.

- **Incomplete Cholesky with threshold**: set the fill-in entries to zeros when they are less than some chosen-*threshold*. It takes the advantage of previous analysis that shows the locations of fill-in entries according to their magnitudes.

We would like to see how good the Incomplete Cholesky with threshold can be applied to find the approximate solution.
**Remark**

Let \( u, U^N \) be the solutions of the continuous and discrete problems respectively. The numerical method yields the discretization error

\[
\| u - U^N \| \leq C_g(N),
\]

where \( \| \cdot \| \) denotes the discrete maximum norm, we wish to find an approximate solution \( U^N_{\text{App}} \) to \( U^N \) in a reasonable way. We have

\[
\| u - U^N_{\text{App}} \| \leq \| u - U^N \| + \| U^N - U^N_{\text{App}} \|.
\]

It is feasible to find \( U^N_{\text{App}} \) such that

\[
\| U^N - U^N_{\text{App}} \| \leq C_g(N).
\]
Let $L_p$ be the Incomplete Cholesky with threshold $\Delta_p = \delta^{2(p+1)} h$, and denote $U_{\text{App}}^{N,p} = (L_p L_p^T)^{-1} b$.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\Delta_p$</th>
<th>$|U^N - U_{\text{App}}^{N,p}|$</th>
<th>$|b - AU_{\text{App}}^{N,p}|$</th>
<th>$|A - L_p L_p^T|$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.342e-08</td>
<td>4.799e-05</td>
<td>1.831e-10</td>
<td>6.984e-16</td>
</tr>
<tr>
<td>3</td>
<td>9.223e-14</td>
<td>5.821e-10</td>
<td>1.776e-15</td>
<td>4.712e-21</td>
</tr>
<tr>
<td>4</td>
<td>2.418e-16</td>
<td>3.492e-10</td>
<td>9.992e-16</td>
<td>8.631e-22</td>
</tr>
<tr>
<td>5</td>
<td>6.338e-19</td>
<td>3.492e-10</td>
<td>9.992e-16</td>
<td>8.570e-22</td>
</tr>
<tr>
<td>6</td>
<td>1.662e-21</td>
<td>3.492e-10</td>
<td>9.992e-16</td>
<td>8.570e-22</td>
</tr>
</tbody>
</table>

**Table:** The Incomplete Cholesky with various thresholds when $\epsilon = 10^{-4}$ and $N = 512.$
Cholesky factorization on a boundary layer-adapted mesh
Boundary layer-adapted meshes for reaction-diffusion problems.

Figure: The Shishkin (left) and Bakhvalov (right) meshes for reaction-diffusion problems.

Note that the nodes in the interior region is similar to that of uniform mesh.
We use the subscripts indicate the block structure of corners, C, edge layers, E and the interior points (not including the nodes along transition points in both directions), I. This gives the partitioned matrix of the form

\[
A = \begin{bmatrix}
A_{CC} & A_{CE} & 0 \\
A_{EC} & A_{EE} & A_{EI} \\
0 & A_{IE} & A_{II}
\end{bmatrix}.
\]

We note that in the interior region where the grids on both x- and y-directions are uniform with the mesh width H is proportional to \(N^{-1}\). Therefore, the matrix \(A_{II}\) shares the same structure like the matrix stencil defined in (3),

\[
A_{II} = \begin{pmatrix}
\mathcal{O}(\varepsilon^2) & \mathcal{O}(\varepsilon^2) \\
\mathcal{O}(\varepsilon^2) & \mathcal{O}(H^2) & \mathcal{O}(\varepsilon^2) \\
\mathcal{O}(\varepsilon^2) & \mathcal{O}(\varepsilon^2)
\end{pmatrix}.
\]

where \(A_{II}\) is an \((N/2 - 1)^2 \times (N/2 - 1)^2\) banded matrix. It inherits all properties of the matrix defined in (3). Therefore, the previous analysis can be applied to the matrix \(A_{II}\).
Again, let $L_p$ be the Incomplete Cholesky with threshold $\Delta_p = \delta^{2(p+1)} h$, and denote $U_{App}^{N,p} = (L_p L_p^T)^{-1} b$.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\Delta_p$</th>
<th>$|U^N - U_{App}^{N,p}|$</th>
<th>$|b - AU_{App}^{N,p}|$</th>
<th>$|A - L_p L_p^T|$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.678e-009</td>
<td>2.028e-005</td>
<td>5.624e-013</td>
<td>1.341e-012</td>
</tr>
<tr>
<td>2</td>
<td>1.100e-012</td>
<td>2.731e-008</td>
<td>8.684e-016</td>
<td>1.328e-015</td>
</tr>
<tr>
<td>5</td>
<td>3.095e-022</td>
<td>2.232e-013</td>
<td>1.084e-019</td>
<td>3.219e-020</td>
</tr>
<tr>
<td>6</td>
<td>2.028e-025</td>
<td>2.232e-013</td>
<td>1.084e-019</td>
<td>3.219e-020</td>
</tr>
</tbody>
</table>

Table: The Incomplete Cholesky with various thresholds when $\epsilon = 10^{-4}$ and $N = 256$ on a Shishkin mesh.

Compared to

$$\|U^N - U_{App}^{N,full}\| = 3.5971 \times 10^{-14},$$
$$\|b - AU_{App}^{N,full}\| = 1.3553 \times 10^{-19},$$
$$\|A - LL^T\| = 1.3559 \times 10^{-20},$$

where $L$ is a complete Cholesky factor of $A$, and $U_{App}^{N,full} = (LL^T)^{-1} b$. 
Conclusion

- We gave a detailed description for the magnitudes of fill-in entries of Cholesky factors for the matrices in which the values of entries have different scales.
- We also provided an estimate for the sum of number of fill-in entries according to their magnitudes.
- The Incomplete Cholesky with threshold shown its advantages for finding the approximate solutions in the context of singularly perturbed problems.


Thank you!