Numerical Challenges in Solving Singularly Perturbed Differential Equations with Layer-Adapted Meshes

Thái Anh Nhan
(nhan@hnu.edu)

Holy Names University, Oakland, CA

Lawrence Berkeley National Lab, September 24 2019
Outline...

1. Intro to SPPs
   - Inadequacy of uniform meshes
   - Layer-adapted mesh and uniform convergence

2. Numerical Analysis for SPPs
   - An open question—boundary layers
   - Special inversion method—interior layers

3. Solvers for SPPs
   - Direct solvers: limitations and its analysis
   - Iterative solvers: difficulties and simple preconditioning approach

4. Conclusion and beyond SPPs...
Collaborators

**Solver and Preconditioning:**
- Niall Madden (NUI Galway, Ireland)
- Scott MacLachlan (Memorial Uni. of Newfoundland, Canada)
- José Luis Gracia (Uni. of Zaragoza, Spain)

**Numerical Analysis:**
- Relja Vulanović (Kent State Uni., Ohio)
- Martin Stynes (NUI Cork, Ireland and Beijing Comput. Science Research Center)
- Torsten Linß (Hagen, Germany)
Two examples

**Example 1 (semiconductor):** The “continuity equation” for electrons\(^1\) in a steady-state scaled model of a one-dimensional semiconductor - with several simplifying assumptions - is

\[
\frac{d^2 n}{dx^2} - \frac{d}{dx} \left[ n \frac{d}{dx} (\psi + \log n) \right] = 0, \tag{1}
\]

\(n\): the electron concentration; \(\psi\): the electrostatic potential; and \(d\psi/dx\) typically very large (perhaps \(10^5\)) on part of its domain.

**Example 2 (the Navier-Stokes eq.):** Consider the following time-dependent Navier-Stokes problem\(^2\) in two space variables \(x\) and \(y\):

\[
\frac{\partial u}{\partial t} - \frac{1}{Re} \Delta u + (u \cdot \nabla) u = -\nabla p, \quad y > 0,
\]

\[\nabla \cdot u = 0, \quad y > 0,
\]

\[u = 0, \quad y = 0, \tag{2}\]

at large Reynolds number \(Re\).


Two examples

The coefficients of the diffusion terms in these two examples are dominated by the convection term coefficients; that is convection-diffusion problems. The first example can be rewritten as

\[- \epsilon u'' + bu' + cu = f, \quad \text{on } (0, 1), \quad u(0) = u(1) = 0, \quad (3)\]

whereas the linearization of (2) yeilds

\[\frac{\partial u}{\partial t} - \epsilon \Delta u + b \cdot \nabla u + cu = f, \quad (4)\]

where \( \epsilon = \frac{1}{\Re} \).

- \( \epsilon \): the perturbation parameter, which tends to zero.
- \( b \equiv 0 \) and \( c \not\equiv 0 \), we have reaction-diffusion problems.

We call the problems like (3) and (4) the singularly perturbed problems (SPPs).

**Goal** is to construct a numerical method which is parameter-robust (also known as “uniformly convergent” and “\( \epsilon \)-uniform”).
Singularly perturbed problems

History:

- Analysis of SPPs began early in the 20th century; approximate solutions constructed from asymptotic expansions;
- Numerical approximations began in the 1970s with the pioneer and seminal work of Bakhvalov\(^3\) and has flourished since the 1990s from the contribution of G. Shishkin\(^4\).

---


Continuous solutions and boundary layers

**Reaction-diffusion example:**

\[-\varepsilon^2 \Delta u + u = e^x + y, \quad (x, y) \in \Omega =: (0, 1)^2, \quad u(\partial \Omega) = 0.\]

**Convection-diffusion example:**

\[-\varepsilon \Delta u - (2 + x)u_x - (3 + y^2)u_y + u = e^x + y, \quad (x, y) \in \Omega =: (0, 1)^2, \quad u(\partial \Omega) = 0.\]

*Figure:* Continuous solutions when $\varepsilon = 1$. 
Continuous solutions and boundary layers

**Reaction-diffusion example:**

\[-\varepsilon^2 \Delta u + u = e^x + y, \quad (x, y) \in \Omega = (0, 1)^2, \quad u(\partial \Omega) = 0.\]

**Convection-diffusion example:**

\[-\varepsilon \Delta u - (2 + x)u_x - (3 + y^2)u_y + u = e^x + y, \quad (x, y) \in \Omega = (0, 1)^2, \quad u(\partial \Omega) = 0.\]

**Figure:** Continuous solutions when $\varepsilon = 10^{-1}$. 
Continuous solutions and boundary layers

**Reaction-diffusion example:**

\[-\varepsilon^2 \Delta u + u = e^x + y, \quad (x, y) \in \Omega =: (0, 1)^2, \quad u(\partial \Omega) = 0.\]

**Convection-diffusion example:**

\[-\varepsilon \Delta u - (2 + x) u_x - (3 + y^2) u_y + u = e^x + y, \quad (x, y) \in \Omega =: (0, 1)^2, \quad u(\partial \Omega) = 0.\]

**Figure:** Continuous solutions when $\varepsilon = 10^{-2}$. 
Continuous solutions and boundary layers

**Reaction-diffusion example:**

\[-\varepsilon^2 \Delta u + u = e^x + y, \quad (x, y) \in \Omega =: (0, 1)^2, \quad u(\partial \Omega) = 0.\]

**Convection-diffusion example:**

\[-\varepsilon \Delta u - (2 + x)u_x - (3 + y^2)u_y + u = e^x + y, \quad (x, y) \in \Omega =: (0, 1)^2, \quad u(\partial \Omega) = 0.\]

**Figure:** Continuous solutions when $\varepsilon = 10^{-4}$. 
Consider simple finite-difference schemes applied to one-dimensional reaction-diffusion and convection-diffusion problems:

**Figure**: Finite difference discretizations on a uniform mesh for reaction diffusion (left) and convection diffusion (right), with $\epsilon = 10^{-4}$, $N = 50$. 
Inadequacy of uniform meshes

**Figure:** Finite difference discretizations on a uniform mesh for reaction diffusion (left) and convection diffusion (right), with $\epsilon = 10^{-4}$, $N = 50$. 
Intro to SPPs

Inadequacy of uniform meshes

<table>
<thead>
<tr>
<th>$\varepsilon^2$</th>
<th>N = 16</th>
<th>N = 32</th>
<th>N = 64</th>
<th>N = 128</th>
<th>N = 256</th>
<th>N = 512</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.17e-03</td>
<td>1.55e-03</td>
<td>3.90e-04</td>
<td>9.76e-05</td>
<td>2.44e-05</td>
<td>6.10e-06</td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td>6.36e-02</td>
<td>1.71e-02</td>
<td>4.36e-03</td>
<td>1.10e-03</td>
<td>2.75e-04</td>
<td>6.88e-05</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>5.06e-02</td>
<td>1.52e-01</td>
<td>2.26e-01</td>
<td>1.13e-01</td>
<td>3.15e-02</td>
<td>8.09e-03</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>5.44e-04</td>
<td>2.11e-03</td>
<td>8.22e-03</td>
<td>3.15e-02</td>
<td>1.13e-01</td>
<td>2.32e-01</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>5.44e-06</td>
<td>2.11e-05</td>
<td>8.32e-05</td>
<td>3.30e-04</td>
<td>1.31e-03</td>
<td>5.21e-03</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>5.44e-08</td>
<td>2.11e-07</td>
<td>8.32e-07</td>
<td>3.30e-06</td>
<td>1.32e-05</td>
<td>5.25e-05</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>5.44e-10</td>
<td>2.11e-09</td>
<td>8.32e-09</td>
<td>3.30e-08</td>
<td>1.32e-07</td>
<td>5.25e-07</td>
</tr>
</tbody>
</table>

Table: The errors, $\|u - U^N\|_\infty$, on a uniform mesh.

Let $\bar{U}^N$ be the usual linear interpolant of $U^N$.

<table>
<thead>
<tr>
<th>$\varepsilon^2$</th>
<th>N = 16</th>
<th>N = 32</th>
<th>N = 64</th>
<th>N = 128</th>
<th>N = 256</th>
<th>N = 512</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.86e-03</td>
<td>2.49e-03</td>
<td>6.23e-04</td>
<td>1.56e-04</td>
<td>3.90e-05</td>
<td>9.75e-06</td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td>2.28e-01</td>
<td>7.30e-02</td>
<td>2.05e-02</td>
<td>5.45e-03</td>
<td>1.41e-03</td>
<td>3.59e-04</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>1.44e+00</td>
<td>1.10e+00</td>
<td>7.25e-01</td>
<td>3.45e-01</td>
<td>1.17e-01</td>
<td>3.34e-02</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>1.52e+00</td>
<td>1.51e+00</td>
<td>1.51e+00</td>
<td>1.46e+00</td>
<td>1.24e+00</td>
<td>8.50e-01</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>1.52e+00</td>
<td>1.51e+00</td>
<td>1.51e+00</td>
<td>1.50e+00</td>
<td>1.50e+00</td>
<td>1.50e+00</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>1.52e+00</td>
<td>1.51e+00</td>
<td>1.51e+00</td>
<td>1.50e+00</td>
<td>1.50e+00</td>
<td>1.50e+00</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>1.52e+00</td>
<td>1.51e+00</td>
<td>1.51e+00</td>
<td>1.50e+00</td>
<td>1.50e+00</td>
<td>1.50e+00</td>
</tr>
</tbody>
</table>

Table: The errors, $\|u - \bar{U}^N\|_\infty$, on a uniform mesh.

Finite difference discretization on a uniform mesh does not converge.
Numerical challenges

Finding numerical solutions to the singularly perturbed problems is a great challenge.

- That is because derivatives of $u$ of order $p$ have magnitude $O(\varepsilon^{-p})$.
- For example, the derivative estimate for one-dimensional convection-diffusion problems

$$|u^{(p)}(x)| \leq C \left\{ 1 + \varepsilon^{-p} e^{-\beta x/\varepsilon} \right\} \text{ for } p = 0, 1, \ldots$$

- Consistency errors (in max norm) is unbounded when $\varepsilon \to 0$, behave like $O(\varepsilon^{-1}N^{-1})$: Stability + Consistency $\Rightarrow$ Convergence does not work!

<table>
<thead>
<tr>
<th>$\varepsilon$</th>
<th>$N = 32$</th>
<th>$N = 64$</th>
<th>$N = 128$</th>
<th>$N = 256$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-1}$</td>
<td>1.56e−02</td>
<td>4.66e−03</td>
<td>1.73e−03</td>
<td>8.64e−04</td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td>4.52e−01</td>
<td>2.09e−01</td>
<td>8.33e−02</td>
<td>3.00e−02</td>
</tr>
<tr>
<td>$10^{-3}$</td>
<td>4.46e+00</td>
<td>2.05e+00</td>
<td>8.18e−01</td>
<td>2.94e−01</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>4.46e+01</td>
<td>2.05e+01</td>
<td>8.16e+00</td>
<td>2.93e+00</td>
</tr>
<tr>
<td>$10^{-5}$</td>
<td>4.46e+02</td>
<td>2.05e+02</td>
<td>8.16e+01</td>
<td>2.93e+01</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>4.46e+03</td>
<td>2.05e+03</td>
<td>8.16e+02</td>
<td>2.93e+02</td>
</tr>
<tr>
<td>$10^{-7}$</td>
<td>4.46e+04</td>
<td>2.05e+04</td>
<td>8.16e+03</td>
<td>2.93e+03</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>4.46e+05</td>
<td>2.05e+05</td>
<td>8.16e+04</td>
<td>2.93e+04</td>
</tr>
</tbody>
</table>

Figure: Consistency errors in max norm⁵.

---

Layer-adapted mesh in 1D Bakhvalov mesh

Given N points, want $O(N)$ in each boundary layer

**Bakhvalov:**
- Equidistant points in interior
- Gradually increase concentration near endpoints
- Distribution given by a smooth mesh function
- Transition point implicitly defined.

**Shishkin:**
- Equidistant points in interior
- Equidistant points near boundary
- Distribution given by a piece-wise differentiable mesh function
- Transition point explicitly defined.
- Slightly worse error bounds than Bakhvalov, but simpler in both theory and practice.

*Figure:* Bahkvalov mesh (top) and Shishkin (below) mesh in 1D.
One can prove

$$\max_{i=0,\ldots,N} |u(x_i) - U_N(x_i)| \leq C \left\{ \begin{array}{ll}
N^{-p} & \text{for a Bakhvalov-type mesh,} \\
(N^{-1} \ln N)^p & \text{for a Shishkin mesh,}
\end{array} \right.$$ 

<table>
<thead>
<tr>
<th>$\epsilon^2$</th>
<th>$N = 16$</th>
<th>$N = 32$</th>
<th>$N = 64$</th>
<th>$N = 128$</th>
<th>$N = 256$</th>
<th>$N = 512$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1$</td>
<td>6.17e-03</td>
<td>1.55e-03</td>
<td>3.90e-04</td>
<td>9.76e-05</td>
<td>2.44e-05</td>
<td>6.10e-06</td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td>6.36e-02</td>
<td>1.71e-02</td>
<td>4.36e-03</td>
<td>1.10e-03</td>
<td>2.75e-04</td>
<td>6.88e-05</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>9.04e-02</td>
<td>3.76e-02</td>
<td>1.44e-02</td>
<td>5.00e-03</td>
<td>1.65e-03</td>
<td>5.23e-04</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>9.08e-02</td>
<td>3.82e-02</td>
<td>1.47e-02</td>
<td>5.11e-03</td>
<td>1.68e-03</td>
<td>5.35e-04</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>9.08e-02</td>
<td>3.83e-02</td>
<td>1.47e-02</td>
<td>5.12e-03</td>
<td>1.69e-03</td>
<td>5.37e-04</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>9.08e-02</td>
<td>3.83e-02</td>
<td>1.47e-02</td>
<td>5.12e-03</td>
<td>1.69e-03</td>
<td>5.37e-04</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>9.08e-02</td>
<td>3.83e-02</td>
<td>1.47e-02</td>
<td>5.12e-03</td>
<td>1.69e-03</td>
<td>5.37e-04</td>
</tr>
</tbody>
</table>

**Table:** The errors, $\|u - U_N\|_\infty$, on a Shishkin mesh.

<table>
<thead>
<tr>
<th>$\epsilon^2$</th>
<th>$N = 16$</th>
<th>$N = 32$</th>
<th>$N = 64$</th>
<th>$N = 128$</th>
<th>$N = 256$</th>
<th>$N = 512$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1$</td>
<td>6.17e-03</td>
<td>1.55e-03</td>
<td>3.90e-04</td>
<td>9.76e-05</td>
<td>2.44e-05</td>
<td>6.10e-06</td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td>3.94e-02</td>
<td>1.00e-02</td>
<td>2.56e-03</td>
<td>6.39e-04</td>
<td>1.60e-04</td>
<td>4.00e-05</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>3.41e-02</td>
<td>9.56e-03</td>
<td>2.44e-03</td>
<td>6.13e-04</td>
<td>1.54e-04</td>
<td>3.84e-05</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>3.42e-02</td>
<td>9.72e-03</td>
<td>2.49e-03</td>
<td>6.26e-04</td>
<td>1.57e-04</td>
<td>3.94e-05</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>3.41e-02</td>
<td>9.72e-03</td>
<td>2.49e-03</td>
<td>6.28e-04</td>
<td>1.58e-04</td>
<td>3.95e-05</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>3.41e-02</td>
<td>9.72e-03</td>
<td>2.49e-03</td>
<td>6.28e-04</td>
<td>1.58e-04</td>
<td>3.95e-05</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>3.41e-02</td>
<td>9.72e-03</td>
<td>2.49e-03</td>
<td>6.28e-04</td>
<td>1.58e-04</td>
<td>3.95e-05</td>
</tr>
</tbody>
</table>

**Table:** The errors, $\|u - U_N\|_\infty$, on a Bakhvalov mesh.
One can prove

$$\max_{i=0,\ldots,N} |u(x_i) - U^N(x_i)| \leq C \begin{cases} N^{-p} \\ (N^{-1} \ln N)^p \end{cases}$$

for a Bakhvalov-type mesh,

for a Shishkin mesh,

<table>
<thead>
<tr>
<th>$\varepsilon^2$</th>
<th>$N = 16$</th>
<th>$N = 32$</th>
<th>$N = 64$</th>
<th>$N = 128$</th>
<th>$N = 256$</th>
<th>$N = 512$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.17e-03</td>
<td>1.55e-03</td>
<td>3.90e-04</td>
<td>9.76e-05</td>
<td>2.44e-05</td>
<td>6.10e-06</td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td>6.36e-02</td>
<td>1.71e-02</td>
<td>4.36e-03</td>
<td>1.10e-03</td>
<td>2.75e-04</td>
<td>6.88e-05</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>9.04e-02</td>
<td>3.76e-02</td>
<td>1.44e-02</td>
<td>5.00e-03</td>
<td>1.65e-03</td>
<td>5.23e-04</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>9.08e-02</td>
<td>3.82e-02</td>
<td>1.47e-02</td>
<td>5.11e-03</td>
<td>1.68e-03</td>
<td>5.35e-04</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>9.08e-02</td>
<td>3.83e-02</td>
<td>1.47e-02</td>
<td>5.12e-03</td>
<td>1.69e-03</td>
<td>5.37e-04</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>9.08e-02</td>
<td>3.83e-02</td>
<td>1.47e-02</td>
<td>5.12e-03</td>
<td>1.69e-03</td>
<td>5.37e-04</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>9.08e-02</td>
<td>3.83e-02</td>
<td>1.47e-02</td>
<td>5.12e-03</td>
<td>1.69e-03</td>
<td>5.37e-04</td>
</tr>
</tbody>
</table>

Table: The errors, $||u - U^N||_\infty$, on a Shishkin mesh.

<table>
<thead>
<tr>
<th>$\varepsilon^2$</th>
<th>$N = 16$</th>
<th>$N = 32$</th>
<th>$N = 64$</th>
<th>$N = 128$</th>
<th>$N = 256$</th>
<th>$N = 512$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.17e-03</td>
<td>1.55e-03</td>
<td>3.90e-04</td>
<td>9.76e-05</td>
<td>2.44e-05</td>
<td>6.10e-06</td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td>3.94e-02</td>
<td>1.00e-02</td>
<td>2.56e-03</td>
<td>6.39e-04</td>
<td>1.60e-04</td>
<td>4.00e-05</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>3.41e-02</td>
<td>9.56e-03</td>
<td>2.44e-03</td>
<td>6.13e-04</td>
<td>1.54e-04</td>
<td>3.84e-05</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>3.42e-02</td>
<td>9.72e-03</td>
<td>2.49e-03</td>
<td>6.26e-04</td>
<td>1.57e-04</td>
<td>3.94e-05</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>3.41e-02</td>
<td>9.72e-03</td>
<td>2.49e-03</td>
<td>6.28e-04</td>
<td>1.58e-04</td>
<td>3.95e-05</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>3.41e-02</td>
<td>9.72e-03</td>
<td>2.49e-03</td>
<td>6.28e-04</td>
<td>1.58e-04</td>
<td>3.95e-05</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>3.41e-02</td>
<td>9.72e-03</td>
<td>2.49e-03</td>
<td>6.28e-04</td>
<td>1.58e-04</td>
<td>3.95e-05</td>
</tr>
</tbody>
</table>

Table: The errors, $||u - U^N||_\infty$, on a Bakhvalov mesh.
One can prove

$$\max_{i=0, \ldots, N} |u(x_i) - U^N(x_i)| \leq C \begin{cases} N^{-p} \\
(N^{-1} \ln N)^p \end{cases}$$

for a Bakhvalov-type mesh,

for a Shishkin mesh,

<table>
<thead>
<tr>
<th>$\varepsilon^2$</th>
<th>$N = 16$</th>
<th>$N = 32$</th>
<th>$N = 64$</th>
<th>$N = 128$</th>
<th>$N = 256$</th>
<th>$N = 512$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1$</td>
<td>6.17e-03</td>
<td>1.55e-03</td>
<td>3.90e-04</td>
<td>9.76e-05</td>
<td>2.44e-05</td>
<td>6.10e-06</td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td>6.36e-02</td>
<td>1.71e-02</td>
<td>4.36e-03</td>
<td>1.10e-03</td>
<td>2.75e-04</td>
<td>6.88e-05</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>9.04e-02</td>
<td>3.76e-02</td>
<td>1.44e-02</td>
<td>5.00e-03</td>
<td>1.65e-03</td>
<td>5.23e-04</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>9.08e-02</td>
<td>3.82e-02</td>
<td>1.47e-02</td>
<td>5.11e-03</td>
<td>1.68e-03</td>
<td>5.35e-04</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>9.08e-02</td>
<td>3.83e-02</td>
<td>1.47e-02</td>
<td>5.12e-03</td>
<td>1.69e-03</td>
<td>5.37e-04</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>9.08e-02</td>
<td>3.83e-02</td>
<td>1.47e-02</td>
<td>5.12e-03</td>
<td>1.69e-03</td>
<td>5.37e-04</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>9.08e-02</td>
<td>3.83e-02</td>
<td>1.47e-02</td>
<td>5.12e-03</td>
<td>1.69e-03</td>
<td>5.37e-04</td>
</tr>
</tbody>
</table>

**Table:** The errors, $||u - U^N||_\infty$, on a Shishkin mesh.

<table>
<thead>
<tr>
<th>$\varepsilon^2$</th>
<th>$N = 16$</th>
<th>$N = 32$</th>
<th>$N = 64$</th>
<th>$N = 128$</th>
<th>$N = 256$</th>
<th>$N = 512$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1$</td>
<td>6.17e-03</td>
<td>1.55e-03</td>
<td>3.90e-04</td>
<td>9.76e-05</td>
<td>2.44e-05</td>
<td>6.10e-06</td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td>3.94e-02</td>
<td>1.00e-02</td>
<td>2.56e-03</td>
<td>6.39e-04</td>
<td>1.60e-04</td>
<td>4.00e-05</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>3.41e-02</td>
<td>9.56e-03</td>
<td>2.44e-03</td>
<td>6.13e-04</td>
<td>1.54e-04</td>
<td>3.84e-05</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>3.42e-02</td>
<td>9.72e-03</td>
<td>2.49e-03</td>
<td>6.26e-04</td>
<td>1.57e-04</td>
<td>3.94e-05</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>3.41e-02</td>
<td>9.72e-03</td>
<td>2.49e-03</td>
<td>6.28e-04</td>
<td>1.58e-04</td>
<td>3.95e-05</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>3.41e-02</td>
<td>9.72e-03</td>
<td>2.49e-03</td>
<td>6.28e-04</td>
<td>1.58e-04</td>
<td>3.95e-05</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>3.41e-02</td>
<td>9.72e-03</td>
<td>2.49e-03</td>
<td>6.28e-04</td>
<td>1.58e-04</td>
<td>3.95e-05</td>
</tr>
</tbody>
</table>

**Table:** The errors, $||u - U^N||_\infty$, on a Bakhvalov mesh.
Short pause...
Hans-Görg Roos and Martin Stynes*

Some open questions in the numerical analysis of singularly perturbed differential equations

1 Introduction

In the last decade, a large number of papers dealing with the numerical analysis of singularly perturbed differential equations have appeared in the research literature. A search of the MathSciNet database for papers published in the years 2005–2014 with MSC Primary Classification 65 (viz., Numerical Analysis) and the phrase “singular* perturb*” [in MathSciNet asterisks are wildcards] yields 879 published works. An overview of this body of work is given in the monograph [78], in Linß’s book [56] on layer-adapted meshes, and in Roos’s survey article [69].
Open question: 2DCD on a B-type mesh

Clearly there is a very healthy level of research activity in this area. But regrettably, the “new” results in many recent papers are merely minor extensions and/or syntheses of older results. (Or worse, they are results that were already known!) Perhaps this is an indication that our area of numerical analysis has reached a mature stage in its development?
Open question: 2DCD on a B-type mesh

Our main focus is the convection-diffusion problem

\[ Lu := -\varepsilon \Delta u - b \cdot \nabla u + cu = f \quad \text{in } \Omega \subset \mathbb{R}^n, \]

\[ u = 0 \quad \text{on } \partial \Omega, \]

(1a)

(1b)

**Question 6.** For an upwind finite difference method applied on a B-type mesh to the convection-diffusion problem (1) with \( n = 2 \), can one prove a discrete maximum norm convergence result

\[ \max_{i,j} |u(x_i, y_j) - u_{ij}^N| \leq CN^{-1} \]

under reasonable hypotheses on the data (e.g., if \( \Omega = (0, 1)^2 \), assume that (5) is valid for \( k = 3 \))? Here \( N \) mesh intervals are used in each coordinate direction, and \( u_{ij}^N \) denotes the computed solution at the point \((x_i, y_j)\).
Open question: 2DCD on a B-type mesh

Linß remarks that

We are not aware of any results for B-type meshes that make use of this truncation error and barrier function technique.

in the survey

Layer-adapted meshes for convection–diffusion problems

Torsten Linß

Institut für Numerische Mathematik, Technische Universität Dresden, D-01062 Dresden, Germany

Received 21 October 2002
Open question: 2DCD on a B-type mesh

**Challenges:** we need

- new barrier function approach\(^6\);
- that works for a B-type mesh even in 1D and extend it to 2D\(^7\).

**Sketch of proof ideas:**

\[ u = s + y. \]

Figure: Bahkvalo mesh (top) and Shishkin (below) mesh in 1D.

- Solution decomposition
- Estimate consistency error; \textbf{S meshes:} nodes in \([0, \tau]\) \textbf{vs B meshes:} all nodes in \([0, 1]\);
- Construct appropriate barrier functions;
- Apply the comparison principle.

\(^7\)N. & Vulanović. Analysis of the truncation error and barrier-functions technique for a Bakhvalov-type mesh, ETNA, 2019.
Short pause...
Consider a singularly perturbed quasilinear boundary-value problem:

$$-\varepsilon \frac{d^2 u}{dx^2} - b(u) \frac{du}{dx} + c(u) = 0, \quad x \in (0, 1), \quad u(0) = A, \quad u(1) = B. \quad (5)$$

Technical assumptions:

- $\frac{d}{du} c(u) \geq c_* > 0, \quad u \in \mathbb{R}$.
- $c(0) = 0, \quad A < 0 < B$.

Then, the problem (5) has a unique solution $u_\varepsilon$, which is strictly monotonically increasing. **Difficulties** arise when $u_\varepsilon$ has one or more interior layers:

- Locations of interior layers are not known in general. But as $\varepsilon \to 0$, layers are approximately located around the point $x_*$, where the solution of the reduced problem is discontinuous (has a shock).
- The corresponding interior layer of the discrete problem is shifted; analogous situation can be observed from numerical solutions of the KdV equation\(^8\).

Interior layer problems

The simplest example is the Lagerstrom-Cole model problem,

\[
-\varepsilon \frac{d^2u}{dx^2} - u \frac{du}{dx} + u = 0, \quad x \in (0, 1), \quad u(0) = A, \quad u(1) = B,
\]

with appropriate conditions on \(A\) and \(B\).

Figure: Asymptotic solution of equation (6) with \(A = -1/2, B = 1\); shock at \(x_* = (1 - A - B)/2 = 1/4\).
Inadequacy of Direct Discretization

We discretize the corresponding conservation form,

$$-\varepsilon \frac{d^2 u}{dx^2} - \frac{d}{dx} a(u) + c(u) = 0, \quad x \in (0, 1), \quad u(0) = A, \quad u(1) = B,$$

where

$$a(u) = \int_0^u b(t)dt,$$

using the Engquist-Osher scheme. It gives

$$-\varepsilon D''u_i - D^- a^-(u_i) - D^+ a^+(u_i) + c(u_i) = 0, \quad i = 1, 2, \ldots, N - 1, \quad (7)$$

where \(u_0 := A, \ u_N := B,\) and

$$D''u_i := \frac{1}{h_i} \left( \frac{u_{i+1} - u_i}{h_{i+1}} - \frac{u_i - u_{i-1}}{h_i} \right),$$

$$D^-u_i := \frac{u_i - u_{i-1}}{h_i}, \quad D^+u_i := \frac{u_{i+1} - u_i}{h_i}.$$

where

$$a^\pm(u) = \int_0^u b^\pm(t)dt, \quad b^+ = \frac{1}{2}(b + |b|), \quad \text{and} \quad b^- = \frac{1}{2}(b - |b|). \quad (8)$$
Inadequacy of Direct Discretization

**Figure:** Numerical solution (and a zoomed-in portion on the right) of Lagerstrom-Cole problem with $A = -1/2, B = 1$, discretized on the Shishkin mesh dense around $x_\ast$. 
Inadequacy of Direct Discretization

**Figure:** Shishkin mesh (left) and uniform mesh (right).

Direct discretization cannot resolve the layer.
Inadequacy of Direct Discretization

Satisfactory results can only be obtained when both continuous and numerical solutions are centrally symmetric with respect to the point \((1/2, 0)\), like when \(-A = B = 1\), giving \(x^* = 1/2\). Figure below shows this situation and illustrates what is meant by a “well-resolved layer”.

Figure: Numerical solution of Lagerstrom-Cole problem with \(A = -1, B = 1\), discretized on the Shishkin mesh.
Inversion Method

Our approach\(^9\) (with Vulanović, 2017) is based on the following observation:

![Graph 1](image1)

![Graph 2](image2)

Figure: Interior layers from a different view.

Inversion Method

We interchange the variables $x$ and $u$ in the original problem. This results in the "inverted problem",

$$
\varepsilon \left( \frac{1}{x'} \right)' - c(u)x' + b(u) = 0, \quad u \in (A, B), \quad x(A) = 0, \quad x(B) = 1,
$$

where $' = d/du$. We discretize the inverted problem on a uniform mesh by a midpoint scheme:

$$
T^N x_i := \varepsilon \left( \frac{1}{\Delta^+ x_i} - \frac{1}{\Delta^- x_i} \right) - D'[c]x_i = -\hat{b}_i, \quad i = 1, 2, \ldots, N - 1,
$$

where

$$
D'[c]x_i := \frac{1 - s_i}{2} \cdot \frac{c_{i-1/2}}{h} \Delta^- x_i + \frac{1 + s_i}{2} \cdot \frac{c_{i+1/2}}{h} \Delta^+ x_i
$$

and

$$
\hat{b}_i = \frac{1 - s_i}{2} b_{i-1/2} + \frac{1 + s_i}{2} b_{i+1/2}
$$

with

$$
s_i = \text{sign } c_i = \begin{cases} 
1 & \text{if } c_i > 0, \\
0 & \text{if } c_i = 0, \\
-1 & \text{if } c_i < 0.
\end{cases}
$$
Inversion Method: Numerical Results

The first test problem is the linear boundary-layer problem

\[-\varepsilon u'' - u' + u = 0, \quad x \in (0, 1), \quad u(0) = -1, \quad u(1) = 1.\]  

(10)

The solution is known, and it has a layer in the neighborhood of \( x = 0 \).

Figure: Inversion method for the problem (10) with \( \varepsilon = 10^{-4}, N = 32 \).
Inversion Method: Numerical Results

The first test problem is the linear boundary-layer problem

\[- \varepsilon u'' - u' + u = 0, \quad x \in (0, 1), \quad u(0) = -1, \quad u(1) = 1. \tag{10}\]

The solution is known, and it has a layer in the neighborhood of \(x = 0\).

<table>
<thead>
<tr>
<th>(-\log_{10} \varepsilon)</th>
<th>(N = 32)</th>
<th>(N = 64)</th>
<th>(N = 128)</th>
<th>(N = 256)</th>
<th>(N = 512)</th>
<th>(\text{Iter}_\varepsilon)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.04e-02</td>
<td>1.05e-02</td>
<td>5.34e-03</td>
<td>2.69e-03</td>
<td>1.35e-03</td>
<td>84</td>
</tr>
<tr>
<td>2</td>
<td>2.54e-02</td>
<td>1.33e-02</td>
<td>6.84e-03</td>
<td>3.47e-03</td>
<td>1.75e-03</td>
<td>82</td>
</tr>
<tr>
<td>3</td>
<td>2.70e-02</td>
<td>1.40e-02</td>
<td>7.22e-03</td>
<td>3.69e-03</td>
<td>1.87e-03</td>
<td>74</td>
</tr>
<tr>
<td>4</td>
<td>2.72e-02</td>
<td>1.41e-02</td>
<td>7.27e-03</td>
<td>3.73e-03</td>
<td>1.89e-03</td>
<td>73</td>
</tr>
<tr>
<td>5</td>
<td>2.73e-02</td>
<td>1.41e-02</td>
<td>7.27e-03</td>
<td>3.73e-03</td>
<td>1.89e-03</td>
<td>74</td>
</tr>
<tr>
<td>6</td>
<td>2.73e-02</td>
<td>1.41e-02</td>
<td>7.27e-03</td>
<td>3.74e-03</td>
<td>1.89e-03</td>
<td>77</td>
</tr>
<tr>
<td>(\text{Ord}^N)</td>
<td>0.95</td>
<td>0.96</td>
<td>0.96</td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table: Errors for the linear boundary-layer problem (10).
Inversion Method: Numerical Results

We consider next the Lagerstrom-Cole problem:

\[-\varepsilon u'' - uu' + u = 0, \quad x \in (0, 1), \quad u(0) = A, \quad u(1) = B.\]

In general, when \(B \leq 1\), \(A \geq -1\), and \(B - A > 1\), the shock is at \(x_* = (1 - A - B)/2\) and the asymptotic solution can be given as

\[
\tilde{u}_\varepsilon(x) := \begin{cases} 
  x + A & \text{if } 0 \leq x < x_* + \frac{1}{\theta} \varepsilon \ln \varepsilon, \\
  \theta \tanh \frac{\theta(x - x_*)}{2\varepsilon} & \text{if } |x - x_*| \leq -\frac{1}{\theta} \varepsilon \ln \varepsilon, \\
  x - 1 + B & \text{if } x_* - \frac{1}{\theta} \varepsilon \ln \varepsilon < x \leq 1,
\end{cases}
\]

where

\[\theta = \frac{B - A - 1}{2} > 0.\]
Inversion Method: Numerical Results

We consider next the Lagerstrom-Cole problem:

\[-\varepsilon u'' - uu' + u = 0, \quad x \in (0,1), \quad u(0) = A, \quad u(1) = B.\]

**Figure:** Direct discretization (left) and inversion method (right) for the Lagerstrom-Cole problem with \( A = -\frac{1}{2}, \quad B = 1 \) (\( x_* = \frac{1}{4} \)).
Inversion Method: Numerical Results

We consider next the Lagerstrom-Cole problem:

\[-\varepsilon u'' - uu' + u = 0, \quad x \in (0, 1), \quad u(0) = A, \quad u(1) = B.\]  \hspace{1cm} (11)

<table>
<thead>
<tr>
<th>$-\log_{10} \varepsilon$</th>
<th>$N = 30$</th>
<th>$N = 60$</th>
<th>$N = 120$</th>
<th>$N = 240$</th>
<th>$N = 480$</th>
<th>$\text{Iter}_{\varepsilon}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1.58e-02</td>
<td>9.29e-03</td>
<td>4.91e-03</td>
<td>2.53e-03</td>
<td>1.22e-03</td>
<td>299</td>
</tr>
<tr>
<td>6</td>
<td>1.62e-02</td>
<td>9.53e-03</td>
<td>5.11e-03</td>
<td>2.70e-03</td>
<td>1.40e-03</td>
<td>109</td>
</tr>
<tr>
<td>7</td>
<td>1.63e-02</td>
<td>9.57e-03</td>
<td>5.13e-03</td>
<td>2.72e-03</td>
<td>1.42e-03</td>
<td>105</td>
</tr>
<tr>
<td>8</td>
<td>1.63e-02</td>
<td>9.58e-03</td>
<td>5.14e-03</td>
<td>2.73e-03</td>
<td>1.42e-03</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>1.63e-02</td>
<td>9.58e-03</td>
<td>5.14e-03</td>
<td>2.73e-03</td>
<td>1.42e-03</td>
<td>97</td>
</tr>
</tbody>
</table>

$\overline{\text{Ord}}^{N}$: 0.77  0.90  0.91  0.94

Table: Errors, $\max_i |\tilde{u}_\varepsilon(x_i) - u_i|$, for the Lagerstrom-Cole problem with $A = -\frac{1}{2}$, $B = 1$ ($x_\ast = \frac{1}{4}$).
Inversion Method: Numerical Results

We consider next the Lagerstrom-Cole problem:

\[-\varepsilon u'' - uu' + u = 0, \quad x \in (0, 1), \quad u(0) = A, \quad u(1) = B. \quad (12)\]

<table>
<thead>
<tr>
<th>$-\log_{10} \varepsilon$</th>
<th>N = 30</th>
<th>N = 60</th>
<th>N = 120</th>
<th>N = 240</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.13e-03</td>
<td>6.18e-04</td>
<td>3.22e-04</td>
<td>1.64e-04</td>
</tr>
<tr>
<td>2</td>
<td>3.49e-03</td>
<td>1.98e-03</td>
<td>1.07e-03</td>
<td>5.52e-04</td>
</tr>
<tr>
<td>3</td>
<td>4.48e-03</td>
<td>3.12e-03</td>
<td>1.86e-03</td>
<td>1.02e-03</td>
</tr>
<tr>
<td>4</td>
<td>5.17e-03</td>
<td>3.36e-03</td>
<td>2.11e-03</td>
<td>1.18e-03</td>
</tr>
<tr>
<td>5</td>
<td>5.42e-03</td>
<td>3.53e-03</td>
<td>2.17e-03</td>
<td>1.21e-03</td>
</tr>
<tr>
<td>6</td>
<td>5.50e-03</td>
<td>3.58e-03</td>
<td>2.19e-03</td>
<td>1.22e-03</td>
</tr>
<tr>
<td>7</td>
<td>5.54e-03</td>
<td>3.60e-03</td>
<td>2.19e-03</td>
<td>1.22e-03</td>
</tr>
<tr>
<td>8</td>
<td>5.55e-03</td>
<td>3.60e-03</td>
<td>2.20e-03</td>
<td>1.22e-03</td>
</tr>
<tr>
<td>9</td>
<td>5.55e-03</td>
<td>3.60e-03</td>
<td>2.20e-03</td>
<td>1.23e-03</td>
</tr>
</tbody>
</table>

| $\text{Ord}^{I,N}$       | 0.52         | 0.67         | 0.84         |

Table: Errors (by double-mesh principle) for the Lagerstrom-Cole problem with $A = -\frac{1}{2}$, $B = 1$ ($x_\ast = \frac{1}{4}$).
Short pause...
Linear solvers for SPPs

When we apply a finite difference discretization to the two-dimensional problem on a layer-adapted mesh with $N$ intervals in each direction, it leads to a linear system of equations which must be solved. We write the system as

$$AU^N = F,$$  \hspace{1cm} (13)

where $A$ is a banded, symmetric and positive definite $(N-1)^2 \times (N-1)^2$ matrix.

To solve (13), one can use direct or iterative solvers. Surprisingly, there has been very few studies that consider the issue of solving the linear systems with efficiency that is robust with respect to the perturbation parameter$^{10}$.

---

Limitations of direct solvers

In Table below, we show the time, in seconds, taken to compute the Cholesky factorization of $A$, and $N = 512$, on a single core of AMD Opteron 2427, 2200 MHz processor, using CHOLMOD (supernodal sparse Cholesky factorization and update/downdate)\textsuperscript{11}.

**Shishkin mesh**

<table>
<thead>
<tr>
<th>$\varepsilon$</th>
<th>$10^{-2}$</th>
<th>$10^{-3}$</th>
<th>$10^{-4}$</th>
<th>$10^{-5}$</th>
<th>$10^{-6}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (s)</td>
<td>58.213</td>
<td>447.533</td>
<td>179.540</td>
<td>101.507</td>
<td>73.250</td>
</tr>
</tbody>
</table>

**Uniform mesh**

<table>
<thead>
<tr>
<th>$\varepsilon$</th>
<th>$10^{-2}$</th>
<th>$10^{-3}$</th>
<th>$10^{-4}$</th>
<th>$10^{-5}$</th>
<th>$10^{-6}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (s)</td>
<td>52.633</td>
<td>496.887</td>
<td>175.783</td>
<td>74.547</td>
<td>45.773</td>
</tr>
</tbody>
</table>

\textsuperscript{11}Chen et al., ACM Trans. Math. Software, 2008
Limitations of direct solvers

The source of difficulty is the presence of subnormal numbers in the calculation of the Cholesky factors\textsuperscript{12}.

**Shishkin mesh**

<table>
<thead>
<tr>
<th>$\epsilon$</th>
<th>$10^{-2}$</th>
<th>$10^{-3}$</th>
<th>$10^{-4}$</th>
<th>$10^{-5}$</th>
<th>$10^{-6}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (s)</td>
<td>58.213</td>
<td>447.533</td>
<td>179.540</td>
<td>101.507</td>
<td>73.250</td>
</tr>
<tr>
<td>Nonzeros in $L$</td>
<td>133,240,632</td>
<td>127,533,193</td>
<td>78,091,189</td>
<td>62,082,599</td>
<td>54,497,790</td>
</tr>
<tr>
<td>Subnormals in $L$</td>
<td>28,282</td>
<td>2,648,308</td>
<td>1,669,345</td>
<td>1,079,992</td>
<td>814,291</td>
</tr>
<tr>
<td>Underflow-zeros</td>
<td>192,709</td>
<td>5,900,148</td>
<td>55,342,152</td>
<td>71,350,742</td>
<td>78,935,551</td>
</tr>
</tbody>
</table>

**Uniform mesh**

<table>
<thead>
<tr>
<th>$\epsilon$</th>
<th>$10^{-2}$</th>
<th>$10^{-3}$</th>
<th>$10^{-4}$</th>
<th>$10^{-5}$</th>
<th>$10^{-6}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (s)</td>
<td>52.633</td>
<td>496.887</td>
<td>175.783</td>
<td>74.547</td>
<td>45.773</td>
</tr>
<tr>
<td>Nonzeros in $L$</td>
<td>133,433,341</td>
<td>128,986,606</td>
<td>56,259,631</td>
<td>33,346,351</td>
<td>23,632,381</td>
</tr>
<tr>
<td>Subnormals in $L$</td>
<td>0</td>
<td>1,873,840</td>
<td>2,399,040</td>
<td>1,360,170</td>
<td>948,600</td>
</tr>
<tr>
<td>Underflow-zeros</td>
<td>0</td>
<td>4,446,735</td>
<td>77,173,710</td>
<td>100,086,990</td>
<td>109,800,960</td>
</tr>
</tbody>
</table>

In the IEEE double precision format: \[ \begin{cases} |x| > 2.2 \times 10^{-308}, \\ 5 \times 10^{-324} < |x| \lesssim 2.2 \times 10^{-308}, \\ |x| \lesssim 5 \times 10^{-324}, \end{cases} \]

To demonstrate this, the figure below shows the sparsity pattern of a section of the system matrix $A$ (left) and the Cholesky factor (right) for the case $N = 256$ and $\epsilon = 1$. The subnormal numbers are highlighted in red (there are none).

The system matrix (left) and Cholesky factor (right) for $N = 256$ and $\epsilon = 1$. 
Next we show the corresponding figures with $\epsilon = 10^{-4}$. Again the subnormal numbers are highlighted in red. Moreover, there are many zero entries. Computing these zeros also involves subnormal arithmetic, so they are also expensive.

The system matrix (left) and Cholesky factor (right) for $N = 256$ and $\epsilon = 10^{-4}$. 
To show that the entries in the factor do decay exponentially, we show below the maximum entry in the $k^{th}$ diagonal entry.

First we show the case $\varepsilon = 1$: entries are smaller in the center of the band, but not extremely so.

A semi-log plot of the maximum entry in the $k^{th}$ diagonal.
To show that the entries in the factor do decay exponentially, we show below the maximum entry in the $k^{th}$ diagonal entry.

First we show the case $\varepsilon = 1$: entries are smaller in the center of the band, but not extremely so.

Next we add the results for $\varepsilon = 10^{-4}$.

A semi-log plot of the maximum entry in the $k^{th}$ diagonal.
Analysis of Cholesky factorization

We wish to investigate the Cholesky factorization of system matrices arising from **finite difference** discretization applied to the **two-dimensional reaction-diffusion** problems on a layer-adapted mesh.

But first, we consider the standard central finite difference discretization on **uniform mesh** with $N$ intervals on each direction. The uniform stepsize is denoted by $h = N^{-1}$ and we will suppose that $\epsilon \ll h$, which is the case of interest. Then the discrete matrix $A$ can be written as the following 5-point stencil

$$A = \begin{bmatrix} -\epsilon^2 & -\epsilon^2 & 4\epsilon^2 + h^2 b_{i,j} & -\epsilon^2 \\ -\epsilon^2 & 4\epsilon^2 + h^2 b_{i,j} & -\epsilon^2 \\ -\epsilon^2 & -\epsilon^2 & 0(h^2) & -\epsilon^2 \\ -\epsilon^2 & -\epsilon^2 & -\epsilon^2 \\ \end{bmatrix}, \quad (14)$$

since $(4\epsilon^2 + h^2 b_{i,j}) = O(h^2)$, where we write $f(\cdot) = O(g(\cdot))$ if there exists positive constants $C_0, C_1$ independent of variables such that $C_0 |g(\cdot)| \leq f(\cdot) \leq C_1 |g(\cdot)|$. 

Analysis of Cholesky factorization

**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, $L^{[m]}$ ($m = N - 1$) 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 $l(i, j) = O(l^{[k]})$. 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 $L^{[\infty]}$.

**Goal:** Classify the fill-in entries associated with their magnitude.
Analysis of Cholesky factorization

We have shown that, it can be determined that $L$ has the following block structure:

$$L = \begin{pmatrix} M & P & Q \\ P & Q & P \\ & & \ddots & \ddots \\ & & P & Q \end{pmatrix},$$

where

$$M = \begin{pmatrix} \Theta(h) & \Theta(h) \\ \Theta(\epsilon^2/h) & \Theta(h) \\ & \Theta(\epsilon^2/h) & \Theta(h) \\ & & \ddots & \ddots \\ & & & \Theta(\epsilon^2/h) & \Theta(h) \end{pmatrix}.$$
Analysis of Cholesky factorization

\[
P = \begin{pmatrix}
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\end{pmatrix},
\]

where the entries belong to \( L^{[k]} \) are denoted by \([k]\).
Analysis of Cholesky factorization

We have shown that

\[ l^{[k]} = O \left( \frac{\epsilon^2 (k+1)}{h^{2k+1}} \right) = O \left( \delta^2 (k+1) h \right), \quad \text{for} \quad k = 1, \ldots, m. \]  

Theorem 1

Let \( \delta = \epsilon / h \). The magnitude \( l^{[k]} \) is

The formulation given in (15) tells us that, as observed experimentally, the values of fill-in entries decay exponentially with respect to \( k \). In practice, for a reaction-dominated problem, we usually have \( \epsilon \ll h \). Hence, when \( \epsilon \) 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.

\[ ^{13} \text{T.N. and Niall Madden, } \text{Cholesky factorization of linear systems coming from finite difference approximations of singularly perturbed problems, 2015.} \]
Analysis of Cholesky factorization

Furthermore, we can give the exact number of the fill-in entries associated with their magnitude.

**Theorem 2**

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

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

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

Combining Theorems 1 and 2, we can predict subnormal and underflow numbers:

<table>
<thead>
<tr>
<th>ε</th>
<th>$10^{-4}$</th>
<th>$10^{-5}$</th>
<th>$10^{-6}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subnormals in L</td>
<td>2,399,040</td>
<td>1,360,170</td>
<td>948,600</td>
</tr>
<tr>
<td>Underflow-zeros</td>
<td>77,173,710</td>
<td>100,086,990</td>
<td>109,800,960</td>
</tr>
</tbody>
</table>

Compared with the exact subnormals and underflows:

<table>
<thead>
<tr>
<th>ε</th>
<th>$10^{-4}$</th>
<th>$10^{-5}$</th>
<th>$10^{-6}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (s)</td>
<td>175.783</td>
<td>74.547</td>
<td>45.773</td>
</tr>
<tr>
<td>Nonzeros in L</td>
<td>56,259,631</td>
<td>33,346,351</td>
<td>23,632,381</td>
</tr>
<tr>
<td>Subnormals in L</td>
<td>2,399,040</td>
<td>1,360,170</td>
<td>948,600</td>
</tr>
<tr>
<td>Underflow-zeros</td>
<td>77,173,710</td>
<td>100,086,990</td>
<td>109,800,960</td>
</tr>
</tbody>
</table>
Overview

The importance of robust iterative solvers

- As stated in Roos et al.\textsuperscript{14}: 
  \textit{It is important to realize that iterative solvers, like the underlying discretization, should be robust with respect to the singular perturbation parameter.}

- However, we show that the condition number of the coefficient matrix grows unboundedly when $\varepsilon$ tends to zero, and so unpreconditioned iterative schemes, such as the conjugate gradient algorithm, perform poorly with respect to $\varepsilon$.

- We provide a careful analysis of \textbf{diagonal} and \textbf{incomplete Cholesky} preconditionings, and show that the condition number of the preconditioned linear system is robust w.r.t. the perturbation parameter. We demonstrate numerically the surprising fact that these schemes are more efficient when $\varepsilon$ is small, than when $\varepsilon$ is $O(1)$.

Due to the nature of *layer adapted meshes* as we seen above, the condition number of the coefficient matrix $A$ resulting from finite difference discretization depends badly on perturbed parameter $\varepsilon$.

### Theorem

The coefficient matrix $A$ arising from the symmetrised finite-difference scheme on the Shishkin mesh satisfies

$$
\kappa_2(A) := \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} = \|A\|_2\|A^{-1}\|_2 \leq C \frac{1}{\varepsilon^2 \ln^2 N},
$$

where $C$ is a constant independent of $\varepsilon$ and $N$.

The numerical result is shown on the table below.

<table>
<thead>
<tr>
<th>$\varepsilon^2$</th>
<th>$N = 16$</th>
<th>$N = 32$</th>
<th>$N = 64$</th>
<th>$N = 128$</th>
<th>$N = 256$</th>
<th>$N = 512$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-4}$</td>
<td>2.42e+02</td>
<td>2.65e+02</td>
<td>6.21e+02</td>
<td>1.56e+03</td>
<td>4.17e+03</td>
<td>1.16e+04</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>2.50e+04</td>
<td>1.80e+04</td>
<td>1.33e+04</td>
<td>1.69e+04</td>
<td>4.56e+04</td>
<td>1.29e+05</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>2.51e+06</td>
<td>1.82e+06</td>
<td>1.34e+06</td>
<td>1.01e+06</td>
<td>7.85e+05</td>
<td>1.30e+06</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>2.52e+08</td>
<td>1.82e+08</td>
<td>1.34e+08</td>
<td>1.01e+08</td>
<td>7.85e+07</td>
<td>6.25e+07</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>2.52e+10</td>
<td>1.82e+10</td>
<td>1.34e+10</td>
<td>1.01e+10</td>
<td>7.86e+09</td>
<td>6.25e+09</td>
</tr>
</tbody>
</table>

**Table:** $\kappa_2(A)$ for the finite difference discretization on a Shishkin mesh.
As a result, as $\varepsilon$ decreases, unpreconditioned CG needs more iterations to achieve a reasonable numerical result. See the results shown on Table below:

<table>
<thead>
<tr>
<th>$\varepsilon^2$</th>
<th>$N = 16$</th>
<th>$N = 32$</th>
<th>$N = 64$</th>
<th>$N = 128$</th>
<th>$N = 256$</th>
<th>$N = 512$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-4}$</td>
<td>25</td>
<td>34</td>
<td>65</td>
<td>122</td>
<td>235</td>
<td>452</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>59</td>
<td>129</td>
<td>223</td>
<td>361</td>
<td>716</td>
<td>1441</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>73</td>
<td>209</td>
<td>578</td>
<td>1315</td>
<td>2276</td>
<td>4149</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>86</td>
<td>287</td>
<td>879</td>
<td>2588</td>
<td>6826</td>
<td>14884</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>93</td>
<td>330</td>
<td>1106</td>
<td>3636</td>
<td>11343</td>
<td>33386</td>
</tr>
</tbody>
</table>

**Table**: Iteration counts for unpreconditioned CG.
Diagonal preconditioner

One of the simplest preconditioning strategies is diagonal preconditioning. Let $D$ be the diagonal matrix in which the diagonal entries are exactly those of the original matrix $A$. Let $A_D = D^{-1/2}AD^{-1/2}$ be the preconditioned matrix. We have shown that the condition number

$$\kappa_2(A_D) \leq C \frac{N^2}{\ln^2 N}.$$ 

<table>
<thead>
<tr>
<th>$\varepsilon^2$</th>
<th>$N = 16$</th>
<th>$N = 32$</th>
<th>$N = 64$</th>
<th>$N = 128$</th>
<th>$N = 256$</th>
<th>$N = 512$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-4}$</td>
<td>1.42e+01</td>
<td>3.95e+01</td>
<td>1.15e+02</td>
<td>3.54e+02</td>
<td>1.10e+03</td>
<td>3.49e+03</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>1.38e+01</td>
<td>3.84e+01</td>
<td>1.12e+02</td>
<td>3.39e+02</td>
<td>1.44e+03</td>
<td>6.12e+03</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>1.37e+01</td>
<td>3.82e+01</td>
<td>1.11e+02</td>
<td>3.34e+02</td>
<td>1.04e+03</td>
<td>3.79e+03</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>1.37e+01</td>
<td>3.82e+01</td>
<td>1.11e+02</td>
<td>3.33e+02</td>
<td>1.03e+03</td>
<td>3.29e+03</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>1.37e+01</td>
<td>3.82e+01</td>
<td>1.11e+02</td>
<td>3.32e+02</td>
<td>1.03e+03</td>
<td>3.28e+03</td>
</tr>
</tbody>
</table>

**Table:** $\kappa_2(A_D)$. 

### Diagonal preconditioner

<table>
<thead>
<tr>
<th>$\varepsilon^2$</th>
<th>$N = 16$</th>
<th>$N = 32$</th>
<th>$N = 64$</th>
<th>$N = 128$</th>
<th>$N = 256$</th>
<th>$N = 512$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-4}$</td>
<td>25</td>
<td>34</td>
<td>65</td>
<td>122</td>
<td>235</td>
<td>452</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>59</td>
<td>129</td>
<td>223</td>
<td>361</td>
<td>716</td>
<td>1441</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>73</td>
<td>209</td>
<td>578</td>
<td>1315</td>
<td>2276</td>
<td>4149</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>86</td>
<td>287</td>
<td>879</td>
<td>2588</td>
<td>6826</td>
<td>14884</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>93</td>
<td>330</td>
<td>1106</td>
<td>3636</td>
<td>11343</td>
<td>33386</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\varepsilon^2$</th>
<th>$N = 16$</th>
<th>$N = 32$</th>
<th>$N = 64$</th>
<th>$N = 128$</th>
<th>$N = 256$</th>
<th>$N = 512$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-4}$</td>
<td>11</td>
<td>19</td>
<td>36</td>
<td>71</td>
<td>140</td>
<td>265</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>11</td>
<td>19</td>
<td>35</td>
<td>69</td>
<td>136</td>
<td>258</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>11</td>
<td>19</td>
<td>35</td>
<td>69</td>
<td>135</td>
<td>266</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>11</td>
<td>19</td>
<td>35</td>
<td>69</td>
<td>147</td>
<td>287</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>11</td>
<td>19</td>
<td>35</td>
<td>69</td>
<td>147</td>
<td>288</td>
</tr>
</tbody>
</table>

**Table**: Numbers of iterations of unpreconditioned CG (top) and the diagonal preconditioned CG (bottom).
Incomplete Cholesky preconditioner

We now choose the preconditioner, \( M = LL^T \) from IC(0). Conditions of being a good preconditioner are satisfied: the system \( Mx = b \) can be easily solved by back-substitution, and that \( M \) is a good approximation to \( A \), because\(^{16}\).

**Theorem**

Let \( Ax = b \) be the linear system for the finite difference discretization on a Shishkin mesh. Let \( M \) be the IC(0) preconditioner. If \( A = M - R \), then

\[
\|R\|_\infty \leq C\varepsilon^2.
\]

We note that in the case where the problem is singularly perturbed, i.e., when \( \varepsilon \to 0 \), then \( \|R\|_2 \leq \|R\|_\infty \to 0 \). Therefore, \( M \) is very close to \( A \), and so the first condition of a good preconditioner is satisfied.

\(^{16}\)T.N. and Niall Madden, *An analysis of simple preconditioners on a layer-adapted mesh*, 2015.
Theorem

Let $M$ define from the IC(0) factorization with $A = M - R$. Denote $A_M = (M^{-1/2}AM^{-1/2})$ a symmetric positive definite matrix. Then

$$\kappa_2(A_M) \leq C \frac{N^2}{\ln^2 N}.$$ 

Sketch of the proof: based on the following theorems:

Theorem: Meijerink and Van Der Vorst (1977)

If $A$ is a symmetric $M$-matrix and $M$ is its IC(0) factorization, then $A = M - R$ is a regular splitting, i.e., $M$ is nonsingular with $M^{-1} \geq 0$ and $M \geq A$ (inequalities should be understood component-wise).

Theorem: R. S. Varga (1962)

Let $A = M - N$ be a regular splitting of $A$, where $A^{-1} \geq 0$. Then

$$\rho(M^{-1}N) < 1.$$
As suggested by the theorem, the condition number is now independent of $\varepsilon$:

\[
\begin{array}{c|ccccccc}
\varepsilon^2 & N = 16 & N = 32 & N = 64 & N = 128 & N = 256 & N = 512 \\
\hline
10^{-4} & 2.42e+02 & 2.65e+02 & 6.21e+02 & 1.56e+03 & 4.17e+03 & 1.16e+04 \\
10^{-6} & 2.50e+04 & 1.80e+04 & 1.33e+04 & 1.69e+04 & 4.56e+04 & 1.29e+05 \\
10^{-8} & 2.51e+06 & 1.82e+06 & 1.34e+06 & 1.01e+06 & 7.85e+05 & 1.30e+06 \\
10^{-10} & 2.52e+08 & 1.82e+08 & 1.34e+08 & 1.01e+08 & 7.85e+07 & 6.25e+07 \\
10^{-12} & 2.52e+10 & 1.82e+10 & 1.34e+10 & 1.01e+10 & 7.86e+09 & 6.25e+09 \\
\end{array}
\]

\[
\begin{array}{c|ccccccc}
\varepsilon^2 & N = 16 & N = 32 & N = 64 & N = 128 & N = 256 & N = 512 \\
\hline
10^{-4} & 1.75 & 3.53 & 9.06 & 29.74 & 106.99 & 361.05 \\
10^{-6} & 1.73 & 3.48 & 8.88 & 25.94 & 80.98 & 262.00 \\
10^{-8} & 1.73 & 3.47 & 8.86 & 25.85 & 80.59 & 260.31 \\
10^{-10} & 1.73 & 3.47 & 8.85 & 25.84 & 80.54 & 260.12 \\
10^{-12} & 1.73 & 3.47 & 8.85 & 25.84 & 80.54 & 260.10 \\
\end{array}
\]

**Table:** $\kappa_2(A)$ (top) and $\kappa_2(A_M)$ (bottom).
Number of iterations: CG vs with the \textbf{IC(0) preconditioned} CG:

<table>
<thead>
<tr>
<th>$\epsilon^2$</th>
<th>$N = 16$</th>
<th>$N = 32$</th>
<th>$N = 64$</th>
<th>$N = 128$</th>
<th>$N = 256$</th>
<th>$N = 512$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-4}$</td>
<td>25</td>
<td>34</td>
<td>65</td>
<td>122</td>
<td>235</td>
<td>452</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>59</td>
<td>129</td>
<td>223</td>
<td>361</td>
<td>716</td>
<td>1441</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>73</td>
<td>209</td>
<td>578</td>
<td>1315</td>
<td>2276</td>
<td>4149</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>86</td>
<td>287</td>
<td>879</td>
<td>2588</td>
<td>6826</td>
<td>14884</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>93</td>
<td>330</td>
<td>1106</td>
<td>3636</td>
<td>11343</td>
<td>33386</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\epsilon^2$</th>
<th>$N = 16$</th>
<th>$N = 32$</th>
<th>$N = 64$</th>
<th>$N = 128$</th>
<th>$N = 256$</th>
<th>$N = 512$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-4}$</td>
<td>3</td>
<td>5</td>
<td>10</td>
<td>19</td>
<td>41</td>
<td>83</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td>17</td>
<td>34</td>
<td>69</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td>17</td>
<td>34</td>
<td>67</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td>17</td>
<td>34</td>
<td>67</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td>17</td>
<td>34</td>
<td>67</td>
</tr>
</tbody>
</table>

\textbf{Table:} Numbers of iterations of unpreconditioned CG (top) and the IC(0)-PCG (bottom).
Solvers for SPPs

Iterative solvers: difficulties and simple preconditioning approach

Solver Time

![Graph showing solver time for different preconditioning approaches](image)

- **Unpreconditioned**
- **Diagonal**
- **IC(0)**
Conclusion

- Robust numerical solutions to SPPs are difficult to achieve, especially to guarantee high accuracy near the layers, even for 1D problems.
- In two dimensions, one should use layer-adapted meshes, but many open remain\(^{17}\).
- The study of solvers for SPPs in the context of specially structured grids is relatively new.

I hope that my talk encourages more computational scientists (probably from CASC?) to enter the fascinating world of singularly perturbed problems!

Thank you for your attention!

---

Conclusion

- Robust numerical solutions to SPPs are difficult to achieve, especially to guarantee high accuracy near the layers, even for 1D problems.
- In two dimensions, one should use layer-adapted meshes, but many open remain\(^{17}\).
- The study of solvers for SPPs in the context of specially structured grids is relatively new.

I hope that my talk encourages more computational scientists (probably from CASC?) to enter the fascinating world of singularly perturbed problems!

Thank you for your attention!

---

Conclusion

- Robust numerical solutions to SPPs are difficult to achieve, especially to guarantee high accuracy near the layers, even for 1D problems.
- In two dimensions, one should use layer-adapted meshes, but many open remain\(^{17}\).
- The study of solvers for SPPs in the context of specially structured grids is relatively new.

I hope that my talk encourages more computational scientists (probably from CASC?) to enter the fascinating world of singularly perturbed problems!

Thank you for your attention!

\(^{17}\text{Roos and Stynes, }\textit{Some open questions in the numerical analysis of singularly perturbed differential equations, }\textit{Computational Methods in Applied Mathematics, }2015\)