DEFECT CORRECTION FOR THE SOLUTION OF A SINGULAR PERTURBATION PROBLEM
AFDELING NUMERIEKE WISKUNDE
(DEPARTMENT OF NUMERICAL MATHEMATICS)

P.W. HEMKER & P.M. DE ZEEUW

DEFECT CORRECTION FOR THE SOLUTION OF A
SINGULAR PERTURBATION PROBLEM

Preprint

kruislaan 413  1098 SJ  amsterdam
Printed at the Mathematical Centre, 413 Kruislaan, Amsterdam.

The Mathematical Centre, founded the 11th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).
Defect correction for the solution of a singular perturbation problem *)

by

P.W. Hemker & P.M. de Zeeuw

ABSTRACT

A method is described for the accurate discretization of a differential equation in which the highest derivative is multiplied by a small parameter. It is well known that for such singular perturbation problems with a strongly asymmetric differential operator almost all discretizations are either unstable or inaccurate or direction dependent. By the combination, in an iterative process, of an inaccurate stable and an accurate unstable scheme we obtain an accurate stable solution, without adapting the scheme to the flow direction. In fact, two approximate solutions are obtained, that—uniformly in ε—are both $O(h^2)$ accurate in the smooth part of the solution. The difference between both solutions can be used for the detection of the unsmooth parts.

KEY WORDS & PHRASES: singular perturbation problem; defect correction; stiff boundary-value problem

*) This report will be submitted for publication elsewhere.
1. INTRODUCTION

An iterative process is used to obtain the accurate solution of a singular perturbation problem. As a model problem we use the convection diffusion equation

\[(1a) \quad -\Delta u + \alpha \cdot \nu u = f,\]
on a bounded domain $\Omega \subset \mathbb{R}^2$, with either Dirichlet or natural boundary conditions

\[(1b) \quad u = g \text{ on } \partial \Omega_1,\]
\[(1c) \quad \nu \cdot \nu u = h \text{ on } \partial \Omega_2,\]

where $\nu$ is the outward normal on the boundary $\partial \Omega = \partial \Omega_1 \cup \partial \Omega_2$.

The problem is written in symbolic form as

\[(2) \quad L \epsilon u = f.\]

It is well known that for such a problem with a strongly asymmetric differential operator, the usual discretizations are either unstable (the usual symmetric discretization methods: central differences, finite elements or Bubnov-Galerkin methods) or inaccurate (artificial viscosity) or direction dependent (various streamline-upwind or Petrov-Galerkin discretizations). By the combination, in an iterative process, of an inaccurate stable and an accurate unstable scheme, we construct a solution which is accurate, uniformly for all $\epsilon$, without adapting the scheme to the subcharacteristic directions (flow-directions) in the problem.

In the iterative process two standard discretizations of (2) are used:

\[(3) \quad L_{\epsilon,h} u_h = f_h,\]
a standard accurate discretization (e.g. standard central differences or a finite element discretization with piecewise linear test and trial functions on a regular triangulation), which is unstable for $\epsilon = h$, and an artificial diffusion (artificial viscosity) discretization

\[(4) \quad L_{\alpha,h} u_h = f_h,\]

which is the same as (3) but with $\alpha = \epsilon + O(h)$, whence (4) is stable. It is well known [6] that both discretizations yield bad results for small ratios $\epsilon/h$.

We combine the discretizations (3) and (4) in an iterative process of defect correction type. In case of a linear problem an elementary defect correction process generates a sequence of approximate solutions $u^{(i)}$, $i = 1,2,3,\ldots$, by the iteration (iterative refinement)

\[(5) \quad u^{(i+1)} = f - L u^{(i)} + \tilde{L} u^{(i)}.\]

If $\lim_{i \to \infty} u^{(i)} = u^*$ and $L$ is injective, then $u^*$ is the solution of the "target equation"

$\text{L} u = f$.

$\tilde{L}$ is usually some approximation to $L$, for which the equation (5) is readily solved.

For our problem we study a "mixed" defect correction iteration with two target operators and two approximate operators. They are combined as:

\[(6a) \quad \tilde{L}_{1,h} u^{(i+1)} = f_1 = L_{1,h} u^{(i)} + \tilde{L}_{1,h} u^{(i)},\]
\[(6b) \quad \tilde{L}_{2,h} u^{(i+1)} = f_2 = L_{2,h} u^{(i+1)} + \tilde{L}_{2,h} u^{(i+1)}.\]

If this iteration converges, we obtain two "solutions" $u^A = \lim_{i \to \infty} u^{(i)}$ and $u^B = \lim_{i \to \infty} u^{(i+1)}$. With $f_1 = f_2 = f$, obviously $u^A$ is characterized by

\[(7) \quad [\tilde{L}_{2,h} - (L_{2,h} - \tilde{L}_{2,h}) \tilde{L}_{1,h}^{-1} (L_{1,h} - \tilde{L}_{1,h})] u^A = [I - (L_{2,h} - \tilde{L}_{2,h}) \tilde{L}_{1,h}^{-1}] f\]
and $u^B$ is given by a similar equation.
2. THE DISCRETIZATION

To obtain an approximate solution for (2), we use the process (6) with \( f_1 = f_2 = f_h \) and with the operators

\[
\begin{align*}
L_1 &= L_{e,h}, \\
L_2 &= L_{a,h}, \\
L_1 &= L_{a,h}, \\
L_2 &= 2 \text{ diag } (L_{a,h}) =: D.
\end{align*}
\]

In this way the second iteration step (6b) is a damped Jacobi-relaxation-step for the solution of (6). The first step (6a) is an order improving defect correction step towards the solution of (3), where the defect defining operator is simply given by

\[
(9) \quad E := L_1 - L_2 = (\alpha - e) A_h.
\]

The two solutions \( u_h^A \) and \( u_h^B \) are now characterized by

\[
\begin{align*}
(10) \quad (L_{e,h} + L_{a,h} D^{-1} E) u_h^A &= f_h, \\
(11) \quad (L_{e,h} + E D^{-1} L_{a,h} ) u_h^B &= (I + E D^{-1}) f_h.
\end{align*}
\]

By using local mode analysis [4] we can show that, for the smooth parts of the solution, \( u_h^A \) and \( u_h^B \) are both accurate of \( O(h^2) \). Further, the operators \( L_{e,h} + L_{a,h} D^{-1} E \) and \( L_{e,h} + E D^{-1} L_{a,h} \) are stable and the numerical boundary layers extend over a region of only \( O(h) \) in the boundary layer region [5].

Remark.

It is not possible to find an accurate approximation for our problem (1) by application of the simple defect correction (5) alone. If we apply (5) with \( L = L_1 \) and \( E = L_2 \), for \( e \) with iteration would not converge to a sensible solution because of the instability of \( L_1 \). It would converge to the unwanted solution of the "target problem" (3). Theoretically, already a single step or (a few steps) of (5) result in a 2nd order accurate method cf. [2] Satz 2.2., from which follows

\[
\| u_h - R_h u \| \leq C h^2 \| u'' \|.
\]

Here \( R_h \) denotes the restriction of the true solution \( u \) to the mesh. However this is not a useful errorbound in our case, where \( \| u'' \| \) may be very large. In fact, for small \( e \), we do not find the \( O(h^2) \) convergence in practice for any reasonable value of \( h \). This is shown in table 1.

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\epsilon &= 10^{-6} & u_h^1 & u_h^2 & u_h^A & u_h^B \\
\hline
h &= 1/8 & 0.790 & 0.634 & 0.608 & 0.459 \\
\text{ratio} &= & 1.37 & 1.73 & 3.82 & 3.46 \\
& & & & & \\
& &= 0.578 & 0.360 & 0.159 & 0.132 \\
\text{ratio} & & 1.62 & 2.02 & 4.75 & 4.64 \\
& & & & & \\
& &= 0.380 & 0.173 & 0.0335 & 0.0291 \\
\text{ratio} & & & & & \\
\hline
\end{array}
\]

Table 1.

Table 1 shows the error in the smooth part of the solution in the \( l_{\infty} \)-norm.

\( u_h^1 \) : the solution after one step (6a);

\( u_h^2 \) : the solution after two steps (6a);

\( u_h^A \), \( u_h^B \) obtained by iteration of (6).

The problem \( \cos u + u = f \) on the unit square;

with the Dirichlet boundary data and the data \( f \) such that

\[
\begin{align*}
\cos x y = & \sin x y \sin y + \cos y x \cos (3 y) + \\
& + (\exp (x / \epsilon) - \exp (-x / \epsilon)) / (1 - \exp (-1 / \epsilon)).
\end{align*}
\]

3. RELATION WITH MULTIGRID TECHNIQUES

The algorithm in section 2 can be considered independent of multigrid techniques. However, it is related to previous work done on multigrid methods.

The use of defect correction in combination with a multigrid algorithm is already mentioned by Brandt[1] and is theoretically studied byHackbusch [2,3]. They consider a multigrid algorithm which, in the elementary form of a two-level algorithm, can be described by a process (6) as well. Then, (6a) is a "coarse grid correction" which is now written in the form

\[
\begin{align*}
L_1 u^{(1)} &= f_h + L_2^{-1} (f_1 - u_h^{(1)}),
\end{align*}
\]

because \( L_1 \) is of deficient rank;

\[
\begin{align*}
L_1 &= P_h 2 h^2 A_{h} 2 h, R_{h} \text{h}, \\
L_1 &= P_h 2 h^2 A_{h} 2 h, R_{h} \text{h},
\end{align*}
\]

where \( P_h \) and \( R_h \) are the grid transfer operators between the fine and the coarse grid; (6b) denotes a sequence of relaxation sweeps.

In the standard two-level algorithm both target operators are the same: \( L = L_1 \). In combination with defect correction (non-standard), the operator \( L_1 \) corresponds to a more accurate discretization than \( L_2 \).

In the approach we discuss here, we differ from [1,2,3] by using a full rank operator \( L_1 = L_{a,h} \).

We use the multigrid algorithm only to solve efficiently the equation (6a).

4. THE DIFFERENCE BETWEEN \( u_h^A \) AND \( u_h^B \)

Considering the difference \( u_h^B - u_h^A \) of the solutions of (10) and (11), we find the following:

\[
\begin{array}{|c|c|c|c|c|}
\hline
\epsilon &= 10^{-6} & u_h^1 & u_h^2 & u_h^A & u_h^B \\
\hline
h &= 1/8 & 0.630 & 0.0740 & 0.0780 & 0.0693 \\
\text{ratio} &= & 2.47 & 3.65 & 3.64 & 3.46 \\
& & & & & \\
& &= 0.0255 & 0.0203 & 0.0214 & 0.0201 \\
\text{ratio} & & 1.71 & 4.02 & 4.01 & 3.89 \\
& & & & & \\
& &= 0.0149 & 0.00503 & 0.00533 & 0.00516 \\
\text{ratio} & & & & & \\
\hline
\end{array}
\]

Table 1a.
Lemma
(12) \( u^B_h - u^A_h = D^{-1}L_{\Delta_t} A \cdot u^{\Delta_t} (L_{\Delta_t})^{-1} A \cdot u_h \).

Proof
Because of (tb) we have
(13) \( D (u^B_h - u^A_h) = L_{\Delta_t} u^B_h - f_h \),
and hence, by (11),
(14) \( L_{\Delta_t} u^B_h - f_h = E D^{-1} (f_h - L_{\Delta_t} u^A_h) =
\) \( (L_{\Delta_t} - L_{\Delta_t}) (u^B_h - u^A_h) \),
or
(15) \( L_{\Delta_t} u^B_h - f_h = E u^A_h \).

From (13) and (15) the lemma follows immediately.

From (12) we see that the difference \( u^B_h - u^A_h \) is large for large \( \Delta u^A_h \). In the regions where \( u^B_h - u^A_h \) is large, the 2nd order differences of \( u^A_h \) are large. Here the solution cannot accurately be represented. Thus, we may use the difference \( u^B_h - u^A_h \) as an indicator where the mesh should be refined for a better approximation of the solution.

To study the local behaviour of the difference \( u^B_h - u^A_h \) in more detail, we resort to the local mode analysis [1]. We consider the problem (1a) on \( \mathbb{R}^2 \) and form the discrete operator \( L_{\Delta_t} \) over all \( \mathbb{R}^2 \). As a forcing function we take the "mode"
\[ f_h(j) = e^{ij \omega}; j \in \mathbb{Z}; h \in \mathbb{R}^2; \text{Re}(\omega) \in [-\pi, \pi]^2 \]

Now the solutions \( u^A_h \) and \( u^B_h \) are studied by Fourier analysis. We note that the Fourier transform is a norm-preserving bijection between the \( \ell_2(\mathbb{Z}^2) \) and the \( L_2((-\pi, \pi)^2) \) functions, i.e.
\[ \|u\| = \|\hat{u}\| \]
where \( \hat{u} \) is the Fourier transform of \( u \).

From (12) we derive
(16) \( u^B_h - u^A_h = FT(\{a-c\} \cdot \text{diag}^{-1}(L_{\Delta_t} A) u^A_h) \cdot \text{diag}(L_{\Delta_t} A) u^A_h) \cdot \text{diag}^{-1}(L_{\Delta_t} A) u^A_h \).

From (10) we find
(17) \( \frac{1}{h} a^B_h = -2\omega^2 \left(\sin^2(\omega_h/2) \right) \cdot \frac{1}{2} \omega^2 \left(\sin^2(\omega_h/2) \right) u^A_h = -2\omega^2 \left(\sin^2(\omega_h/2) \right) u^A_h \),
and from (17) we find
(18) \( \frac{1}{h} a^B_h = -\frac{1}{2} \omega^2 \left(\sin^2(\omega_h/2) \right) \cdot \frac{1}{2} \omega^2 \left(\sin^2(\omega_h/2) \right) + \frac{1}{h} T (1+\frac{\omega^2}{h^2}) \),

where \( 4\pi^2/h^2 = 1 \) is the characteristic trigonometric polynomial for the reduced difference operator.

In previous work [4,5] we derived from (17) that \( u_h^A \) is bounded, uniform in \( h \). Moreover, \( u_h^B \) is \( O(h) \) accurate in the smooth parts of the solution. We also showed that, for small \( c \), at boundary or interior layers the error in \( u_h^B \) may be \( O(1) \), uniformly in \( c \). In these regions (where the discrete solution is not able to represent the true solution anyway) the numerical approximations may show oscillations. However, the discretization (10) is asymptotically stable and the critical regions near the layers have only \( O(h) \) width.

To see the effect of \( h = 0 \) on the approximation, we study separately the "low" and the "high" frequencies in the true solution. The frequencies are called "high" or "low" with reference to the mesh used (size \( h \)). For the "low" frequencies in the solution we consider a fixed \( \omega \) and let \( h \to 0 \). For the "high" frequencies we consider \( h \) fixed and let \( h \to 0 \). Regions where the mesh is too coarse to represent a solution properly, are dominated by high frequencies.

Now we use (16) to see how \( u^B_h - u^A_h \) behaves in the different regions. For the low frequencies \( S^2 = \sin^2(\omega_h/2) + \sin^2(\omega_h/2) = O(h^2) \) and we find
\[ \| u^B_h - u^A_h \| \leq \frac{\| \omega^2 \|}{2\omega^2} \| u^A_h \| \leq \frac{\| \omega^2 \|}{2\omega^2} \| u^A_h \| \]
Because \( u^A_h \) is bounded, uniform in \( c \), we conclude that in smooth part of the solution we have
\[ \| u^B_h - u^A_h \| \leq C h^2 \]
with \( C \) independent of \( h \).

For the high frequencies \( \omega \) is fixed and hence also \( S^2 \) and we find
\[ \| u^B_h - u^A_h \| \leq \frac{\| \omega^2 \|}{2\omega^2} \| u^A_h \| \leq \frac{\| \omega^2 \|}{2\omega^2} \| u^A_h \| \]
Hence, for \( h \to 0 \), \( C \) fixed, we find
\[ \| u^B_h - u^A_h \| = O(h) \]
for the high frequencies.

Hence we find that in the smooth part of the solution (low frequencies) the difference \( u^B_h - u^A_h \) is small, whereas it may be large in those regions of transition layers, where the mesh is not fit to represent the true solution. In this way the difference \( u^B_h - u^A_h \) behaves similar to the error in the approximation \( u^B_h - u^A_h \).

5. EXAMPLES
As a first example to show the behaviour of \( u^A_h \) and \( u^B_h - u^A_h \), we used the same equation as Hughes and Brooks [6],
\[ -\Delta u + \cos(\theta) u_x + \sin(\theta) u_y = 0 \]
on the unit square, with Dirichlet boundary conditions at the inflow boundary and \( c = 10^{-6} \).

\[ \text{Inflow boundary conditions} \]
\[ \begin{align*}
\text{If } x = 0, & \quad 0 \leq x \leq 1, \\
\text{If } y = 0, & \quad 0 \leq y \leq A, \\
\text{If } y = A, & \quad 0 \leq y > A.
\end{align*} \]
\[ A = 3/16. \]
Fig. 1a. Example 1. Neumann BCs, $\theta = 22.5^\circ$, $u_h^A$.

Fig. 1b. $u_h^B - u_h^A$. Z

Fig. 2a. Example 1. Neumann BCs, $\theta = 22.5^\circ$, $u_h^A$.

Fig. 2b. $u_h^B - u_h^A$. N

Fig. 3a. Example 1. Neumann BCs, $\theta = 45^\circ$, $u_h^A$.

Fig. 3b. $u_h^B - u_h^A$.

Fig. 4a. Example 1. Neumann BCs, $\theta = 67.5^\circ$, $u_h^A$.

Fig. 4b. $u_h^B - u_h^A$. 
Fig. 5a. Example 1. Dirichlet BCs, $\theta = 22.5^\circ$, $u_h^A$.

Fig. 5b. $u_h^B - u_h^A$.

Fig. 6a. Example 1. Dirichlet BCs, $\theta = 45^\circ$, $u_h^A$.

Fig. 6b. $u_h^B - u_h^A$.

Fig. 7a. Example 1. Dirichlet BCs, $\theta = 67.5^\circ$, $u_h^A$.

Fig. 7b. $u_h^B - u_h^A$.

Fig. 8a. Example 2. $u_h^A$.

Fig. 8b. Example 2. $u_h^B - u_h^A$. 
At the outflow boundary either natural or homogeneous Dirichlet boundary conditions are used. For $L_2$ we use the usual finite element discretization with piecewise linear functions on a regular triangulation.

Note: the way the squares are triangulized ($S$ or $Z$) makes no essential difference in the results. If the triangle division follows the internal layer, the numerical results are — of course — slightly better. In the figures the triangulation is indicated.

In common with $[n]$ we use as flow directions $\theta = 22.5^\circ$, $45^\circ$ and $67.5^\circ$.

In the figures (1)-(7) we show the numerical solution $u_h^n$ and the difference $u_h^n - u_A^n$ for the different cases.

As a second example we use the variable coefficient problem:

\[-cu + \Delta u = 0 \quad \text{on } [-1, +1]^2 \times [0,1],\]

\[A = (y(-x^2), -x(-y^2)).\]

with the Dirichlet boundary conditions

\[u(x,y) = 1 + \tanh(10x+20y), \quad y=0, -1 \leq x < 0,
\]

\[u(x,y) = 0, \quad y=1, -1 \leq x \leq 1,
\]

\[u(x,y) = 0, \quad x=0, 0 \leq y \leq 1,
\]

and homogeneous Neumann boundary conditions at the outflow boundary: $y=0, 0 < x < 1$.

Asymptotically for $c \to 0$, the true solution is constant over the subcharacteristics and the outflow profile is the mirror image of the inflow profile.

In the figures (8a) and (8b) we show the numerical solution $u_h^n$ and the difference $u_h^n - u_A^n$.

REFERENCES


