THEORETICAL ANALYSIS OF THE UPWIND FINITE VOLUME SCHEME ON THE COUNTER-EXAMPLE OF PETERSON

DANIEL BOUCHE\textsuperscript{1}, JEAN-MICHEL GHIDAGLIA\textsuperscript{2} AND FRÉDÉRIC P. PASCAL\textsuperscript{2}

Abstract. When applied to the linear advection problem in dimension two, the upwind finite volume method is a non consistent scheme in the finite differences sense but a convergent scheme. According to our previous paper [Bouche et al., SIAM J. Numer. Anal. 43 (2005) 578–603], a sufficient condition in order to complete the mathematical analysis of the finite volume scheme consists in obtaining an estimation of order \(p\), less or equal to one, of a quantity that depends only on the mesh and on the advection velocity and that we called geometric corrector. In [Bouche et al., Hermes Science publishing, London, UK (2005) 225–236], we prove that, on the mesh given by Peterson [SIAM J. Numer. Anal. 28 (1991) 133–140] and for a subtle alignment of the direction of transport parallel to the vertical boundary, the infinite norm of the geometric corrector only behaves like \(h^{1/2}\) where \(h\) is a characteristic size of the mesh. This paper focuses on the case of an oblique incidence i.e. a transport direction that is not parallel to the boundary, still with the Peterson mesh. Using various mathematical technics, we explicitly compute an upper bound of the geometric corrector and we provide a probabilistic interpretation in terms of Markov processes. This bound is proved to be like \(h\), so that the order of convergence is one. Then the reduction of the order of convergence occurs only if the direction of advection is aligned with the boundary.

Mathematics Subject Classification. 65M06, 65M12, 65M15, 76M12.

Received July 8, 2009. Revised December 17, 2009. Published online March 17, 2010.

INTRODUCTION

We consider on the domain \(\Omega \subset \mathbb{R}^2\), the following advection problem with a constant velocity \(a \in \mathbb{R}^2\)

\[
\frac{\partial u}{\partial t} + \nabla \cdot (au) = 0
\]

(0.1)

where \(u\) is defined from \(\mathbb{R}^2 \times \mathbb{R}\) to \(\mathbb{R}\). This problem is completed with initial and boundary data that satisfy the so called compatibility conditions.

The space discretization is based on an unstructured mesh composed of volumes \(K_j\) where \(j\) is the index of the volume and \(V_j\) the surface of the volume. The upwind finite volume method takes into account the direction

$^{1}$ CEAM, DAM, DIF, 91297 Arpajon, France.
$^{2}$ CMLA, ENS Cachan, CNRS, Universit\'e, 61 avenue du Pr\'esident Wilson, 94235 Cachan Cedex, France.
frederic.pascal@cmla.ens-cachan.fr
where the information comes from and consists in approximating on each control volume the mean value of \( u \) denoted by:

\[
v^n_j = \frac{1}{V_j} \int_{K_j} u(x, n \Delta t) \, dx.
\]

If the time discretization is for instance the Euler forward scheme, then the unknowns \( u^n_j \) which approximate \( v^n_j \) satisfy

\[
\mathcal{L}(u_j) = u^n_{j+1} - u^n_j + \frac{\Delta t}{V_j} \sum_{k \in \mathcal{N}^- (j)} a \cdot \mathbf{N}_{jk} (u^n_k - u^n_j) = 0
\]

where

- \( \mathbf{N}_{jk} \) is the normal to the interface \( K_j \cap K_k \) that points out from \( K_j \) to \( K_k \), with a norm equal to the length of the interface \( |K_j \cap K_k| \);
- \( \mathcal{N}^- (j) \) denotes the set of indices of adjacent volumes to \( K_j \) such that \( a \) is inward to \( K_j \) on the common interface \( i.e. \mathcal{N}^- (j) = \{ k, K_k \text{ adjacent to } K_j \text{ and } a \cdot \mathbf{N}_{jk} < 0 \} \).

It turns out that this scheme is not consistent in the sense of finite differences: the truncation error \( \mathcal{L}(v_j) \), which is obtained by substituting \( u^n_j \) by the exact value \( v^n_j \), does not converge to zero as \( h \) goes to zero. Then Lax theorem is helpless, see for instance [7]. However numerous theoretical results (see [3,9,12,15,21,22]) obtained with solutions more or less regular prove the convergence of the scheme. Specifically, the order of convergence is known to be at least \( \frac{2}{3} \) in \( L^2 \) norm for \( H^1 \) data (Desprès [6]) or in \( L^p \) norm for \( W^{1,p} \) and \( BV \) data (Merlet and Vovelle [16] or Delarue and Lagoutière [5]). This property of enhancement of the truncation error is called supra-convergence and was widely analyzed in various cases but different from the finite volume one, in [14,20,23,24].

In [1], the concept of geometric corrector is introduced in order to analyze the scheme from a mathematical point of view. This corrector, for which existence and uniqueness have been proved, depends only on the mesh and on the advection velocity. If \( K_j \) is not on the boundary of the domain, then the corrector \( \Gamma_j \) associated to the volume \( K_j \) is a point in \( \mathbb{R}^2 \) which satisfies the following system (see [1] for the general case)

\[
\sum_{k \in \mathcal{N}^+ (j)} a \cdot \mathbf{N}_{jk} \Gamma_j + \sum_{k \in \mathcal{N}^- (j)} a \cdot \mathbf{N}_{jk} \Gamma_k = \sum_{k \in \mathcal{N}^+ (j)} a \cdot \mathbf{N}_{jk} (g_{jk} - g_j) + \sum_{k \in \mathcal{N}^- (j)} a \cdot \mathbf{N}_{jk} (g_{jk} - g_k)
\]

where

- \( g_j \) denotes a point inside the volume like for instance the center of gravity of the volume \( K_j \);
- \( g_{jk} \) denotes the center of gravity of the edge \( K_j \cap K_k \);
- \( \mathcal{N}^+ (j) \) is the set of indices of adjacent volumes to \( K_j \) such that \( a \) is outward to the common interface \( i.e. \mathcal{N}^+ (j) = \{ k, K_k \text{ adjacent to } K_j \text{ and } a \cdot \mathbf{N}_{jk} > 0 \} \).

For \( 0 < p \leq 1 \), we prove in [1] that, when the solution is regular enough, a \( h^p \) behavior of the norm of the corrector is a sufficient condition for the scheme to be of order \( p \). As a result, estimating the corrector is an efficient way to study the convergence of the scheme.

It was proved in [1] that, in dimension 2, on non structured meshes of triangles, uniformly refined, the standard upwind scheme, although not consistent, is in fact of order one, at least for regular enough solutions. The proof of this result for meshes that are independently refined is still an open problem.

Indeed, slower convergence appears for vertical advection velocity on a particular mesh originally proposed by Peterson [18]. The domain \([0,1] \times [0,1]\) and meshes he considered are plotted on Figure 1. One can observe that there is no vertical line (except the boundaries). Moreover, the refinement is uniform except near vertical boundaries. The index of the triangle \( j \) previously introduced is replaced in the present case, by a pair of integers \((m,n)\) and a sign. The triangles far from the vertical boundaries are rectangle isosceles triangles of length \( h = \frac{1}{m+n} \) and are denoted by \( T^+_{m,n} \) for upward ones and by \( T^-_{m,n} \) for downward ones. Indices \( m \) and \( n \) are of opposite parity. Triangles adjacent to the left boundary are rectangle isosceles triangle of length \( \frac{1}{\sqrt{2}} \) and are denoted by \( T^+_{0,n} \) and \( T^-_{0,n} \) (with \( n \) odd) as it is indicated on Figure 2. Triangles adjacent to the right boundary
The aim of this article is to obtain, by different mathematical technics, estimates on the geometric corrector as a function of mesh size $h$ for oblique incidence, i.e. for a non vertical advection direction, in order to prove the following theorem:

**Theorem 0.1.** For oblique incidence, we have the estimation

$$\|\Gamma\|_\infty \equiv \sup_{m,n} \|\Gamma_{m,n}\| \leq c h.$$ 

Equations satisfied by the corrector are given in Section 1. Since left and right boundary conditions do not interfere, the study is reduced to a semi-infinite domain, obtained by extending to the right the original one. The following sections establish an upper bound of the norm of the corrector successively by counting geometric paths, then by using a recursive argument and finally by Fourier analysis. The upper bound is a partial sum of a convergent series whose sum studied in Section 5 depends on $h$. In Section 6, we shall return to the vertical advection direction case. The use of generating functions allows us to prove in a different way the result proved in [2]: in that case, the series is divergent and the corrector behaves as $h^{1/2}$ for the uniform norm. Finally in Section 7 which deals with the original square, we propose a probabilistic interpretation, similar in spirit to that in the paper by Delarue et Lagoutière [5], and we use a method based on matrices to get the upper bound.
1. Equations satisfied by the corrector $\Gamma_{m,n}$

We denote by $\Gamma^+_{m,n}$ the corrector associated to triangle $T^+_{m,n}$, by $\Gamma^-_{m,n}$ the corrector associated to $T^-_{m,n}$, and define $\theta$ as the angle between the advection velocity and the vertical direction (cf. Fig. 1). The advection velocity incomes on the left boundary, and we suppose that $0 < \theta < \frac{\pi}{4}$. We note

$$\alpha = \frac{\pi}{4} - \theta.$$ 

The corrector is computed, choosing as reference point $g^+_{m,n}$, the middle of the height of the triangle $T^+_{m,n}$ for $m \neq 0, 2\ell$ and the middle of the vertical side of the triangle for $m = 0$ and $m = 2\ell$ as shown in Figure 2. This choice is more convenient than the center of gravity of the triangle. Since we denote by $\vec{A}_{m,n}$ the corrector associated to triangle $T^+_{m,n}$ for $m \neq 0, 2\ell$, namely setting $\vec{A}_{m,n} = 0$ and the middle of the vertical side of the triangle for $m = 0, 2\ell$, can be written, following (0.3)

$$\begin{align*}
\cos \theta (\Gamma^+_{m,n} - \Gamma^-_{m,n}) &= \frac{h_2}{4} \left( \cos \theta \vec{j} + \sqrt{2} \left( \cos \left( \frac{\pi}{4} \right) - \cos \left( \frac{\pi}{4} + \frac{\pi}{4} \right) \right) \vec{j} \right) \\
\cos \theta \Gamma^-_{m,n} &= \frac{\sqrt{2}}{2} \cos \left( \frac{\pi}{4} \right) \Gamma^+_{m-1,n-1} - \frac{\sqrt{2}}{2} \cos \left( \frac{\pi}{4} + \frac{\pi}{4} \right) \Gamma^+_{m+1,n-1} \\
&= \frac{h_2}{4} \left( \cos \theta \vec{j} + \sqrt{2} \left( \cos \left( \frac{\pi}{4} \right) - \cos \left( \frac{\pi}{4} + \frac{\pi}{4} \right) \right) \vec{j} \right).
\end{align*}$$

By adding these equalities, we obtain

$$\Gamma^+_{m,n} = \frac{\cos \alpha}{\cos \alpha + \sin \alpha} \Gamma^+_{m-1,n-1} - \frac{\sin \alpha}{\cos \alpha + \sin \alpha} \Gamma^+_{m+1,n-1} = 0$$

and setting $p = \frac{\cos \alpha}{\cos \alpha + \sin \alpha}$ and $q = \frac{\sin \alpha}{\cos \alpha + \sin \alpha}$, we get the recursion relation

$$\Gamma^+_{m,n} = p\Gamma^+_{m-1,n-1} + q\Gamma^+_{m+1,n-1}. \quad (1.1)$$

We now consider the triangles with their vertical side lying on the left boundary, i.e. $m = 0$. The equations satisfied by the corrector can be written, following (0.3)

$$\begin{align*}
\frac{\sqrt{2}}{2} \cos \alpha \Gamma^+_{0,2k+1} - \frac{1}{2} \cos \theta \Gamma^-_{0,2k+1} &= \frac{\sqrt{2}}{2} \cos \alpha \frac{h_2}{4} \vec{i} - \frac{1}{2} \cos \theta \left( \frac{h_2}{4} \vec{i} + \frac{h_2}{4} \vec{j} \right) \\
\frac{1}{2} \cos \theta \Gamma^-_{0,2k+1} - \frac{\sqrt{2}}{2} \sin \alpha \Gamma^+_{1,2k} &= \frac{\sqrt{2}}{2} \sin \alpha \frac{h_2}{4} \vec{i} + \frac{1}{2} \cos \theta \left( \frac{h_2}{4} \vec{i} + \frac{h_2}{4} \vec{j} \right).
\end{align*}$$

By adding these equalities, we obtain

$$\frac{\sqrt{2}}{2} \cos \alpha \Gamma^+_{0,2k+1} - \frac{\sqrt{2}}{2} \sin \alpha \Gamma^+_{1,2k} = \frac{\sqrt{2}}{2} \cos \alpha \frac{h_2}{4} \vec{i} + \frac{\sqrt{2}}{2} \sin \alpha \frac{h_2}{4} \vec{i}$$

namely, setting $\vec{A} = \frac{1}{4}(1 + \tan \alpha)\vec{i}$, vector independent on $k$ and $r = \frac{\pi}{4}$,

$$\Gamma^+_{0,2k+1} - r\Gamma^+_{1,2k} = \vec{A}h. \quad (1.2)$$

For $m = 2\ell$, namely for the triangles with a side on the left boundary, we obtain

$$\Gamma^+_{2\ell,2k+1} - \Gamma^+_{2\ell-1,2k} = \vec{B}h \quad (1.3)$$
where $\vec{B}$ is a vector independent on $k$ and $r$ defined by

$$\vec{B} = \frac{1}{4} \left( \tan \theta - \frac{3}{2} - \frac{1}{2} \tan \alpha \right) \vec{i} + \frac{1}{8} (\tan \alpha - 1) \vec{j}.$$

Gathering the results above, equations (1.1), (1.2), (1.3) satisfied by the correctors associated with upward directed triangles write

$$\begin{align*}
\Gamma_{m,n}^+ &= p \Gamma_{m-1,n-1}^+ + q \Gamma_{m+1,n-1}^+ & m + n \text{ odd, } 0 < m < 2\ell, 0 < n \\
\Gamma_{0,2k+1}^+ - r \Gamma_{1,2k}^+ &= \vec{A} \hat{h} & 0 \leq k < \ell \\
\Gamma_{2\ell,2k+1}^+ - \Gamma_{2\ell-1,2k}^+ &= \vec{B} \hat{h} & 0 \leq k < \ell \\
\Gamma_{m,0}^+ &= 0 & m \text{ odd, } 0 < m < 2\ell. 
\end{align*}$$

(1.4)

Our assumption on advection velocity direction implies $0 < q < \frac{1}{2} < p < 1$ with $p+q = 1$ and $r = \frac{q}{p} < 1$. The difference equations (1.4) satisfied by the corrector have zero right hand side, except on the boundary. We shall show that the order of convergence of the scheme is determined by accumulation of boundary contributions.

**Remark 1.1.** Note that, for vertical advection velocity, namely $\theta = 0$, we have $\alpha = \frac{\pi}{4}$, $p = q = \frac{1}{2}$, $A = -B = \frac{\vec{i}}{2}$ and the above system reduces to system (27) given in Section 4.2 of [2] namely:

$$\begin{align*}
\Gamma_{m,n}^+ &= \frac{1}{2} \Gamma_{m-1,n-1}^+ + \frac{1}{2} \Gamma_{m+1,n-1}^+ \\
\Gamma_{0,2k+1}^+ &= \Gamma_{1,2k}^+ + \frac{h}{2} \vec{i} \\
\Gamma_{2\ell,2k+1}^+ - \Gamma_{2\ell-1,2k}^+ &= \frac{h}{2} \vec{i} \\
\Gamma_{m,0}^+ &= 0. 
\end{align*}$$

(1.5)

**Remark 1.2.** For the case $\theta = \frac{\pi}{4}$, the recursion for the correctors boils down to

$$\Gamma_{m,n}^+ = \Gamma_{m-1,n-1}^+$$

and noting that $\Gamma_{0,2k+1}^+ = \frac{h}{2} \vec{i}$, we obtain $\Gamma_{m,n}^+ = \frac{h}{4} \vec{i}$ for $m \leq n$ and 0 otherwise.

**Remark 1.3.** The case of advection velocity incoming on the right boundary can be deduced by symmetry.

**Remark 1.4.** The first equation of system (1.4) is valid for $m$ and $n$ of opposite parities, i.e. for upward triangles. In order to simplify the computations and the presentation, it is convenient to introduce fictitious triangles $T_{m,n}^+$ and associated correctors $\Gamma_{m,n}^+$ for $m$ and $n$ with same parities. The equation (1.1) can then be extended to $m$ and $n$ of arbitrary parity and we obtain

$$\begin{align*}
\Gamma_{m,n}^+ &= p \Gamma_{m-1,n-1}^+ + q \Gamma_{m+1,n-1}^+ & 0 < m < 2\ell, 0 < n \\
\Gamma_{0,n}^+ - r \Gamma_{1,n-1}^+ &= \vec{A} \hat{h} & 0 < n \\
\Gamma_{2\ell,n}^+ - \Gamma_{2\ell-1,n-1}^+ &= \vec{B} \hat{h} & 0 < n \\
\Gamma_{m,0}^+ &= 0 & 0 < m < 2\ell. 
\end{align*}$$

(1.6)

The two lattices, $m$ and $n$ of opposite parities, and $m$ and $n$ of same parities, are independent from each other, because equations (1.6) relate quantities, either of opposite parities, or of same parities. Equations (1.6) extended to $m$ and $n$ of arbitrary parities therefore allow to determine the correctors associated to “real” triangles, i.e. with $m$ and $n$ of opposite parities.
**Remark 1.5.** To deal with only one boundary, one may consider, as in [2] triangle \([0, 1] \times [0, 1]\), lower left half of the square as shown in Figure 3.

In the following sections, we shall rather consider the semi-infinite mesh defined on the domain \([0, +\infty) \times [0, 1]\): this mesh is obtained by extending Peterson original mesh towards infinity to the right with (rectangular equilateral) triangles with basis \(h\). We therefore have to solve the system

\[
\begin{align*}
\Gamma^+_{m,n} &= p \Gamma^+_{m-1,n-1} + q \Gamma^+_{m+1,n-1} & 0 < m, & 0 < n \\
\Gamma^+_{0,n} &= r \Gamma^+_{1,n-1} + Ah & 0 < n \\
\Gamma^+_{m,0} &= 0 & 0 < m
\end{align*}
\]  

(1.7)

and it is what we do in the following sections.

### 2. Computation by counting geometric paths

The geometric corrector \(\Gamma^+_{m,n}\) is a linear combination of \(\Gamma^+_{m-1,n-1}\) with weight \(p\) and \(\Gamma^+_{m-1,n+1}\) with weight \(q\). Let us represent this relation by two paths, one from point \((m, n)\) to point \((m - 1, n - 1)\) with weight (or probability) \(p\), and one from \((m, n)\) to \((m - 1, n + 1)\) with weight (or probability) \(q\). By recursion, \(\Gamma^+_{m,n}\) is therefore a linear combination of the \(\Gamma^+_{m+t,n-k}\) with weights associated to paths linking point \((m, n)\) to \((m + t, n - k)\) with \(-k \leq t \leq k\). The weight associated to a path composed of \(g\) elementary paths (or displacements) to the left hand side and \(d\) paths to the right hand side is \(p^gq^d\).

If \(m \geq n\), all paths starting from \((m, n)\) reach points on the line \(n = 0\) with coordinates \((m + t, 0)\) with \(-n \leq t \leq n\) before reaching the left boundary. Thus, according to (1.7), \(\Gamma^+_{m,n}\) is equal to zero.

If \(m < n\), some paths starting from \((m, n)\) reach the left boundary. When a path reaches the left boundary at point \((0, n')\), we have

\[
\Gamma^+_{0,n'} = r \Gamma^+_{1,n'-1} + Ah.
\]

Thus, a contribution to the corrector proportional to the weight associated to the path from \((m, n)\) to \((0, n')\) appears. The path is reflected with weight \(r\). As a result, the corrector can be computed as a sum of weights associated with paths issued from point \((m, n)\) and reaching the left boundary. Therefore, to compute the corrector, we simply have to enumerate paths issued from point \((m, n)\) and reaching this boundary.

For arbitrary reflection coefficient \(r\), to the best of our knowledge, this problem has no simple closed form solution.
However, for the specific case $r = 2q$, the problem at hand can be solved by using Désiré André reflection principle [4,19]. To be more specific, we associate bijectively to a path reflected on the left boundary the path obtained by reflection (symmetry) with respect to this boundary. To this end, we introduce the correctors $\Gamma_{m,n}^+$ for $m < 0$, and then split the reflected path with weight $2q$ in two paths with weight $q$ symmetric with respect to the boundary (see Fig. 4). As a result, the relation

$$\Gamma_{0,n}^+ = r\Gamma_{1,n-1}^+ + \vec{A}h$$

is replaced by

$$\Gamma_{0,n}^+ = q\Gamma_{-1,n-1}^+ + q\Gamma_{1,n+1}^+ + \vec{A}h.$$ 

Note that the roles of $p$ and $q$ are swapped but, using this trick, we are able to count the paths from point $(m,n)$ to a point $(0,n-s)$ lying on the left boundary and composed of $s$ elementary displacements among whose $\frac{1}{2}(s+m)$ are towards the left hand side and $\frac{1}{2}(s-m)$ towards the right hand side, with $s$ varying from $m$ (minimum to reach the left boundary) and $n-1$ (the point $(0,0)$ does not contribute). A path is thus defined by the choice of the displacements to the right among the $s$ displacements. The number of possible choices is

$$\left(\frac{s}{2} \right) \left( \frac{s}{2} \right) \left( \frac{s}{2} \right) \left( \frac{s}{2} \right)$$

and we finally get:

$$\Gamma_{m,n}^+ |_{r=2q} = \vec{A}h s \in [m,n-1], s+m=0|2 \sum p^{s+j} q^{s-j} \left( \frac{2s}{s-j} \right), 0 \leq j \leq n.$$ 

For the point $(2j, 2n+1)$, one obtains

$$\Gamma_{2j,2n+1}^+ |_{r=2q} = \vec{A}h \sum_{s=j}^{n} p^{s+j} q^{s-j} \left( \frac{2s}{s-j} \right), 0 \leq j \leq n. \quad (2.1)$$

Since the reflection coefficient $r$ is less than $2q$, $\Gamma_{2j,2n+1}^+ |_{r=2q}$ provides an upper bound for the corrector $\Gamma_{2j,2n+1}^+$.

### 3. Proof by recursion

In this section, we demonstrate the result (2.1) by recursion. Let us suppose that (2.1) is satisfied up to order $n$. Then, for $0 < j \leq n+1$ and for $r = 2q$

$$\Gamma_{2j,2n+3}^+ |_{r=2q} = p^2 \Gamma_{2j-2,2n+1}^+ |_{r=2q} + 2pq \Gamma_{2j,2n+1}^+ |_{r=2q} + q^2 \Gamma_{2j+2,2n+1}^+ |_{r=2q}.$$
And therefore

\[ \Gamma^+_{2j,2n+3|r=2q} = \tilde{A}h \sum_{s=j}^{n+1} p^{s+j}q^{s-j} \left( \frac{2s-2}{s-j} \right) + 2\tilde{A}h \sum_{s=j+2}^{n+1} p^{s+j}q^{s-j} \left( \frac{2s-2}{s-j-2} \right) \]

\[ = \tilde{A}h \left( p^{2j} \left( \frac{2j-2}{0} \right) + p^{2j+1}q \left( \frac{2j}{1} \right) + 2p^{2j+1} \left( \frac{2j}{0} \right) \right) + \tilde{A}h \sum_{s=j+2}^{n+1} p^{s+j}q^{s-j} \left( \frac{2s-2}{s-j} \right) \]

\[ = \tilde{A}h \left( p^{2j} + p^{2j+1}q(2j+2) \right) + \tilde{A}h \sum_{s=j+2}^{n+1} p^{s+j}q^{s-j} \left( \frac{2s}{s-j} \right) \]

Because \( r = 2q \), a similar computation shows that

\[ \Gamma^+_{0,2n+3|r=2q} = \tilde{A}h \sum_{s=0}^{n+1} p^{s+j}q^{s-j} \left( \frac{2s}{s-j} \right). \]

As a result, (2.1) is satisfied up to order \( n + 1 \). Finally, for \( n = 0 \), \( \Gamma^+_{0,1} = \tilde{A}h \) and the recursion relation (2.1) is satisfied for \( n = 0 \), which concludes the proof.

4. PROOF USING FOURIER ANALYSIS

An alternative demonstration of the result is given below. It is inspired from a method studied in the book by Karlin [11] based on the construction of a basis of orthogonal polynomials (in our case related to Tchebytchev polynomials). To this end, we consider \( \Gamma^+_{m,n} \) as the Fourier coefficients of a function \( f_n \). More precisely, we write

\[ \Gamma^+_{m,n} = \int_{-\pi}^{\pi} \left( \frac{p}{q} \right)^m f_n(x) \cos(mx) \, dx. \]  

(4.1)

Recursion relations (1.7) satisfied by the correctors determine \( f_n \). Namely, we get, for any \( m > 0 \)

\[ \int_{-\pi}^{\pi} \left( \frac{p}{q} \right)^m f_{n+1}(x) \cos(mx) \, dx = \int_{-\pi}^{\pi} \left( \frac{p}{q} \right)^m f_n(x) \cos((m-1)x) \, dx \]

\[ + \int_{-\pi}^{\pi} \left( \frac{p}{q} \right)^{m+1} f_n(x) \cos((m+1)x) \, dx \]

which can be written

\[ \int_{-\pi}^{\pi} \left( \frac{p}{q} \right)^m (f_{n+1}(x) - 2\sqrt{pq} \cos x f_n(x)) \cos(mx) \, dx = 0. \]  

(4.2)

It implies that there exists \( c_n \) such that

\[ f_{n+1}(x) = 2\sqrt{pq} \cos x f_n(x) + c_n. \]  

(4.3)

(4.4)
But, for \( m = 0 \),

\[
\int_{-\pi}^{\pi} f_{n+1}(x) \, dx - r \int_{-\pi}^{\pi} \left( \frac{p}{q} \right) \frac{1}{2} f_n(x) \cos x \, dx = \bar{A}h
\]

(4.5)

which can be written, by using relation (4.4)

\[
\int_{-\pi}^{\pi} (2\sqrt{pq} \cos x \, f_n(x) + c_n) \, dx - \int_{-\pi}^{\pi} r \left( \frac{p}{q} \right) \frac{1}{2} f_n(x) \cos x \, dx = \bar{A}h.
\]

(4.6)

If \( r = 2q \), upper bound of \( r = \frac{2}{p} \) for the problem at hand, the equation (4.6) simply provides

\[
c_n = \frac{\bar{A}h}{2\pi}
\]

which, with recursion relation (4.4) for \( f_n \) and \( \Gamma_{m,0}^+ = 0 \) (which implies \( f_0 = 0 \)), gives the expression below

\[
f_{n+1}(x) = \frac{\bar{A}h}{2\pi} \sum_{s=0}^{n} \left( \sqrt{4pq \cos x} \right)^s.
\]

(4.7)

As a result, we get,

\[
\Gamma_{m,n+1}^+|_{r=2q} = \frac{\bar{A}h}{2\pi} \int_{-\pi}^{\pi} \left( \frac{p}{q} \right) \sum_{s=0}^{n} \left( \sqrt{4pq \cos x} \right)^s \cos(mx) \, dx.
\]

(4.8)

Specifically,

\[
\Gamma_{2j,2n+1}^+|_{r=2q} = \frac{\bar{A}h}{2\pi} \left( \frac{p}{q} \right)^j \int_{-\pi}^{\pi} \sum_{s=0}^{2n} \left( \sqrt{4pq \cos x} \right)^s \cos(2jx) \, dx
\]

\[
= \frac{\bar{A}h}{2\pi} \left( \frac{p}{q} \right)^j \sum_{s=0}^{2n} (4pq)^{\frac{s}{2}} \int_{C} 2^{-(s+1)} \left( z + \frac{1}{z} \right)^s \left( z^{2j} + z^{-2j} \right) \, dz
\]

\[
= \frac{\bar{A}h}{2\pi} \left( \frac{p}{q} \right)^j \sum_{s=0}^{2n} (4pq)^{\frac{s}{2}} \int_{C} 2^{-(s+1)} \sum_{k=0}^{s} \binom{s}{k} z^{2k-s} \left( z^{2j} + z^{-2j} \right) \, dz
\]

The only nonzero terms stems from the \( k \) satisfying \( 2k = s \pm 2j \), which on the one hand implies \( s \) even and on the other \( s \geq 2j \) so that \( 0 \leq k \leq s \). Changing variable, we obtain again relation (2.1):

\[
\Gamma_{2j,2n+1}^+|_{r=2q} = \frac{\bar{A}h}{2\pi} \left( \frac{p}{q} \right)^j \sum_{s=2j}^{2n} (4pq)^{\frac{s}{2}} \int_{C} 2^{-s} \left( \frac{s}{(s-2j)/2} \right) \, dz
\]

\[
= \bar{A}h \left( \frac{p}{q} \right)^j \sum_{s=2j}^{n} (pq)^{\frac{s}{2}} \binom{s}{s-j}
\]

\[
= \bar{A}h \sum_{s=j}^{n} p^{j+s} q^{s-j} \binom{2s}{s-j}.
\]
5. Sum of the series and upper bound of the corrector

By using formula 2.5.15 of [25] or the methods provided by [8], the series $S_j(x)$ with term $x^r (2r)^{-j}$ turns out to be summable in closed form. For $0 < x < \frac{1}{4}$, this sum is given by

$$S_j(x) \equiv \sum_{k=j}^{+\infty} x^k \left( \frac{2k}{k-j} \right) = \frac{x^j}{\sqrt{1-4x}} \left( \frac{1 - \sqrt{1-4x}}{2x} \right)^{2j}. \quad (5.1)$$

We deduce from this result the following upper bound for the corrector

$$\|\Gamma_{2j,2n+1}^+\| \leq \|\Gamma_{2j,2n+1}^+\|_{r=2q} < \|A\| h \left( \frac{p}{q} \right)^j S_j(pq).$$

As $p + q = 1$ and $p > q$, we get $\sqrt{1-4pq} = p - q$ and therefore

$$\|\Gamma_{2j,2n+1}^+\| < \frac{\|A\| h}{p-q} = \|A\| \cot \theta. \quad (5.2)$$

We conclude that the norm of the corrector is bounded by $ch$, with $c = \|A\| \cot \theta$. This proves Theorem 0.1.

In the next section, we turn back to the case of vertical advection.

6. Vertical advection

For vertical advection velocity, it is possible to count the number of paths (1.5) with the original (i.e. without taking an upper bound) reflection coefficient:

$$\Gamma_{2j,2n+1}^+ = h \frac{2}{n} \sum_{s=j}^{n} 2^{-2s} \binom{2s}{s-j}. \quad (6.1)$$

By this path counting technic, we obtain the result proven by recursion for a triangular domain in [2].

Another possible method relies on the use of generating functions. We define $\Gamma_{m,n}^+ = \Gamma_{-m,n}^-$ for $m < 0$ and split the paths reflected with weight 1 into two paths of weight $\frac{1}{2}$, symmetric with respect to the left boundary. We have now to solve the equations below, namely

$$\begin{cases} 
\Gamma_{m,n}^+ = \frac{1}{2} \Gamma_{m-1,n-1}^+ + \frac{1}{2} \Gamma_{m+1,n-1}^+ \quad m \in \mathbb{Z}^* \quad n \geq 1 \\
\Gamma_{0,n}^+ = \frac{1}{2} \Gamma_{-1,n-1}^+ + \frac{1}{2} \Gamma_{1,n+1}^+ + \frac{h}{2} \quad n \geq 1 \\
\Gamma_{m,0}^+ = 0 \quad m \in \mathbb{Z}. 
\end{cases} \quad (6.2)$$

The generating function defined by

$$\Gamma_n(z) = \sum_{m \in \mathbb{Z}} \Gamma_{m,n}^+ z^{2m}$$

verifies $\Gamma_0(z) = 0$, $\Gamma_1(z) = \frac{h}{2} z^2$ and the recursion relationship

$$\Gamma_{2n+1}(z) = \frac{1}{4} \left( z + \frac{1}{2} \right)^2 \Gamma_{2n-1}(z) + \frac{h}{2} z^2$$
from which we obtain

\[ \Gamma_{2n+1}(z) = \frac{h}{2} \sum_{k=0}^{n} \left( z + \frac{1}{z} \right)^{2k} \frac{1}{2^{2k}}. \]

Noting that \( \Gamma_{2j,2n+1}^+ \) is the coefficient of \( z^{2j} \) in the above sum, and using the binomial identity, we obtain once again formula (6.1). Indeed, \( \Gamma_{2n+1}(z) = \frac{h}{2} \sum_{k=0}^{n} \sum_{s=0}^{k} \binom{2k}{s} \frac{1}{2^{2k}} \) \( z^{2k-s} z^{-s} \) \( \Gamma_{2n+1}(z) = \frac{h}{2} \sum_{k=0}^{n} \sum_{j=-k}^{k} \left( \binom{2k}{k-j} \frac{1}{2^{2k}} \right) z^{2j}. \)

The series \( S_j(x) \) with term \( x^r \binom{2r}{r-j} \) is divergent for \( x = \frac{1}{4} \). For vertical advection velocity, in contrast to oblique incidence case, the corrector behaves for \( L_\infty \) norm as \( h^{1/2} \) (see [2]). Note that, for this specific case, Peterson in [18] demonstrates that the convergence error of the scheme is exactly \( h^{1/2} \). Then, the corrector provides a fair estimate of the error.

7. BACK TO THE INITIAL DOMAIN: A MATRIX FORMULATION

First, a probabilistic interpretation of the recursion relation (1.6) is proposed, relying on similar problems studied in the realm of Markov process [10]. If we consider a Markov process with discrete states indexed by \( m \) and discrete times indexed by \( n \) and if the square \( 2\ell + 1 \) matrix

\[
T = \begin{pmatrix}
0 & r & 0 & 0 & \ldots & 0 \\
p & 0 & q & 0 & \ldots & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & q & 0 & p \\
0 & 0 & 0 & 0 & r' & 0
\end{pmatrix}
\]

is the transition matrix of Markov process considered above, the transition probability from state \( m \) to state \( m-1 \) is \( p \), whereas this probability from state \( m \) to state \( m+1 \) is \( q \), for \( m \neq 0 \) and \( m \neq 2\ell \). The transition matrix \( T \) allows to compute the \( 2\ell + 1 \) component vector of occupation probabilities \( p(n+1) \) at time \( n+1 \) from the occupation probabilities \( p(n) \) at time \( n \):

\[ p(n+1) = T p(n) = T^{n+1} p(0). \]

The quantity \( \sum_{k=0}^{2n} (T^k)_{j,0} \) provides the sum of occupation probabilities of state 0 at times \( k \leq 2n \) for initial state \( j \), i.e. at time 0, the occupied state is \( j \). The first and last lines of \( T \) respectively dictate the behavior at the left (resp. right) boundary and the coefficients \( r \) and \( r' = 1 \) are the reflection coefficients.

We now express under matrix form the relations (1.6). To this end, we introduce for \( n \geq 0 \), the \( \ell + 1 \) component vector

\[ \Gamma_n = \left( \Gamma_{0,n}^+, \Gamma_{2,n}^+, \ldots, \Gamma_{2\ell,n}^+ \right)^t, \]

and we get

\[ \Gamma_1 = Ch \quad \text{and} \quad \Gamma_{2n+1} = M \Gamma_{2n-1} + Ch, \quad n \geq 1. \]
where $M$ is the square $\ell + 1$ matrix defined by ($r' = 1$)

\[
M = \begin{pmatrix}
rp & rq & 0 & 0 & \ldots & 0 \\
p^2 & 2pq & q^2 & 0 & \ldots & 0 \\
0 & p^2 & 2pq & q^2 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & p^2 & 2pq & q^2 \\
0 & 0 & 0 & 0 & 2p^2 & 2pq \\
\end{pmatrix}
\]

and $C = \begin{pmatrix}
A \\
0 \\
0 \\
\vdots \\
0 \\
\end{pmatrix}$.

Matrix $M$ is the restriction to odd indexes of $T^2$. We deduce from above that

\[
\Gamma_{2n+1} = \sum_{k=0}^{n} M^k Ch.
\]

Therefore, the corrector is a linear combination of $\vec{A}h$ and $\vec{B}h$, with coefficients that are sums of coefficients of powers of $M$:

\[
\Gamma_{2j,2n+1}^+ = a_{j,n}(M)\vec{A}h + b_{j,n}(M)\vec{B}h, \quad 0 \leq j \leq \ell
\]

where

\[
a_{j,n}(M) = \sum_{k=0}^{n} (M^k)_{j,0} \quad \text{and} \quad b_{j,n}(M) = \sum_{k=0}^{n} (M^k)_{j,\ell}
\]

and $(M^k)_{j,i}$ denotes the term on line $j$ and column $i$. The lines and columns of $M^k$ are indexed from 0 to $\ell$.

One will notice that, for even indices, the estimate for the $\ell$ component vector $\Gamma_{2n}^+$ can be obtained from the estimate for vector $\Gamma_{2n-1}$.

To allow closed form computation of eigenvectors, one multiplies by $2p > 1$ the first line of the matrix $M$ in order to replace the terms $rp = q$ and $rq = \frac{q^2}{p}$ respectively by $2pq$ and $2p^2$, which amounts to replace the reflection coefficient $r = \frac{q}{p}$ by an upper bound $2p$ and one multiplies by $2p > 1$ the first line of the matrix $M$, which amounts to replace the reflection coefficient $r' = 1 < 2p$ by an upper bound $2p$. One then obtains the matrix $N$ of order $\ell + 1$:

\[
N = \begin{pmatrix}
2pq & 2q^2 & 0 & 0 & \ldots & 0 \\
p^2 & 2pq & q^2 & 0 & \ldots & 0 \\
0 & p^2 & 2pq & q^2 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 2p^2 & 2pq \\
\end{pmatrix}.
\]

The way $N$ is constructed from $M$ implies that the coefficients $a_{j,n}(N)$ and $b_{j,n}(N)$ computed for this matrix $N$ are upper bounds of those computed with matrix $M$. The $\ell + 1$ eigenvalues of $N$ are distinct and are given by

\[
\lambda_k = 2pq \left( 1 + \cos \left( \frac{k\pi}{\ell} \right) \right) < 1;
\]

the associated eigenvectors have components

\[
X_{jk} = \left( \frac{p}{q} \right)^j \cos \left( \frac{jk\pi}{\ell} \right), \quad 0 \leq j \leq \ell.
\]
In order to compute the coefficients of powers of $N$, we diagonalize this matrix: $N = XDX^{-1}$ where $X$ is the matrix $X_{jk}$ of eigenvectors and $D$ the diagonal matrix of eigenvalues. The inverse of $X$ can be computed, using trigonometric identities given in [13],

$$
\sum_{j=1}^{\ell-1} \cos\left(\frac{n j \pi}{\ell}\right) \cos\left(\frac{m j \pi}{\ell}\right) + \frac{1}{2} (1 + \cos(n\pi) \cos(m\pi)) = \frac{\ell}{2} \delta_{nm}, \quad m \neq 0, \ell, \\
\sum_{j=1}^{\ell-1} \cos\left(\frac{n j \pi}{\ell}\right) \cos\left(\frac{m j \pi}{\ell}\right) + \frac{1}{2} (1 + \cos(n\pi) \cos(m\pi)) = \ell \delta_{nm}, \quad m = 0, \ell.
$$

(7.3)

Defining $y_{mj} = \frac{1}{\ell} (\frac{m}{n}) \cos(\frac{m j \pi}{\ell})$ for $1 \leq j \leq \ell - 1$, $y_{m0} = \frac{1}{2\ell}$, $y_{m\ell} = \frac{1}{2\ell} (\frac{m}{n}) \cos(m\pi)$, then setting $Y_{mj} = 2y_{mj}$ for $m \neq 0, \ell$ and $Y_{m\ell} = y_{m\ell}$ for $m = 0, \ell$, we get the inverse of $X$. As a result, we obtain

$$
(N^k)_{j0} = \sum_{s=0}^{\ell} X_{js} Y_{s0} \lambda_s^k
$$

$$
= \frac{1}{\ell} (2pq)^k \left(\frac{p}{q}\right)^j \left( \sum_{s=1}^{\ell-1} \cos\left(\frac{j s \pi}{\ell}\right) \frac{1}{\ell} \lambda_s^k + \frac{1}{2\ell} \lambda_0^k + \frac{1}{2\ell} \lambda_0^k \right)
$$

$$
= \frac{1}{\ell} (2pq)^k \left(\frac{p}{q}\right)^j \left( \sum_{s=1}^{\ell-1} \cos\left(\frac{j s \pi}{\ell}\right) \left(1 + \cos\left(\frac{s \pi}{\ell}\right)\right)^k + 2^{k-1} \right)
$$

$$
= \frac{1}{\ell} (4pq)^k \left(\frac{p}{q}\right)^j \left( \sum_{s=0}^{\ell} f\left(\frac{s \pi}{\ell}\right) - \frac{1}{2} (f(0) + f(\pi)) \right)
$$

with $f(x) = \cos(jx) \cos^{2k}(\frac{x}{\pi})$. The Euler Mac-Laurin summation formula allows to write this sum as

$$
\sum_{s=0}^{\ell} f\left(\frac{s \pi}{\ell}\right) - \frac{1}{2} (f(0) + f(\pi)) = \frac{\ell}{\pi} \int_0^\pi f(x) \, dx + \frac{\ell}{\pi} \sum_{m=2}^{\infty} (-1)^m \frac{B_{m-1}}{(2m-2)!} (f^{[2m-3]}(\pi) - f^{[2m-3]}(0)).
$$

As $f$ and $f(\cdot + \pi)$ are even functions, all odd order derivatives of $f$ are zero for 0 and $\pi$. Therefore, the sum is equal to the integral $I$ defined below, which is computed by residue

$$
I = \frac{\ell}{\pi} \int_0^\pi f(x) \, dx = \begin{cases} k 2^{-2k} \binom{2k}{k-j} & k \geq j \\\n0 & k < j. \end{cases}
$$

Using this result, we obtain, for $k \geq j$

$$
(N^k)_{j0} = (4pq)^k \left(\frac{p}{q}\right)^j \frac{1}{\ell \pi} \pi 2^{-2k} \binom{2k}{k-j} = p^{k+j} q^{k-j} \binom{2k}{k-j}
$$

and we get

$$
a_{j,n}(N) = \sum_{k=j}^{n} p^{k+j} q^{k-j} \binom{2k}{k-j} = \left(\frac{p}{q}\right)^j \sum_{k=j}^{n} (pq)^k \binom{2k}{k-j}.
$$

By the same method, we obtain

$$
b_{j,n}(N) = \left(\frac{p}{q}\right)^{j-\ell} \sum_{k=\ell-j}^{n} (pq)^k \binom{2k}{k-(\ell-j)}.
$$
Thus, the upper bound of the corrector for the initial square domain turns out to be the sum of the correctors for two semi-infinite meshes computed previously, one obtained by extending to infinity the initial mesh to the right hand side, the other doing the same, but on the left hand side.

8. Conclusion

By different methods (geometric paths counting, recursion, Fourier analysis, generating function, matrix computation), we have established closed form expressions of an upper bound of the geometric corrector for the initial (square) Peterson and related (semi-infinite, triangular) domains for oblique advection velocity. The most convenient expression appears to be a weighted sum of binomial coefficients. This form has a simple probabilistic interpretation, in terms of geometric paths counting. Moreover, the sum of the associated series is known in closed form. As a result, an explicit upper bound for the geometric corrector proportional to $h \cot \theta$ is obtained, where $\theta$ is the angle of the advection velocity with the vertical.

The $L_\infty$ norm of the corrector is therefore of order $h$ for oblique incidence. As a result, the upwind scheme on initial (square) Peterson mesh and on related (triangular, semi-infinite) meshes is therefore of order $h$ for oblique incidence. The reduction of the order of convergence to $h^{1/2}$ for vertical direction of advection in $L_\infty$ norm is a subtle phenomena, which disappears as soon as the advection velocity moves apart from vertical direction, or when norm $L_\infty$ is replaced by norm $L_1$. We have observed the same kind of phenomena in numerical experiments on more generic meshes, see [17]. The main advantage of Peterson mesh is that the difference equations satisfied by the corrector have constant coefficients, and nonzero right hand side only on the boundary, which simplifies the derivation of an upper bound for the corrector. The precise behavior of the corrector on arbitrary meshes remains an open question.

References


