Gas dynamics and Heat and Mass Transfer

Numerical Solution of the Convection–Diffusion Equations

Student: Pedro López Sancha

Professor: Carlos-David Pérez Segarra

Aerospace Technology Engineering
The School of Industrial, Aerospace and Audiovisual Engineering of Terrassa
Technic University of Catalonia

September 11, 2021

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Escola Superior d'Enginyeries Industrial,
Aeroespacial i Audiovisual de Terrassa

Contents

1 Convection–diffusion equations

In this section we derive the continuity equation and the general convection–diffusion equation. To begin, we present and prove Reynolds Transport Theorem, which is a generalization of Leibniz integral rule. Next we deduce the aforementioned equations using this theorem.

1.1 Notation and assumptions

First of all, we shall introduce some notation that will be exhaustively used in the project.

- Let Ω be a subset of $\mathbb{R}^n$. The subsets $\partial\Omega$ and $\overline{\Omega}$ of $\mathbb{R}^n$ will denote the boundary and the closure of Ω, respectively.
- Let $x \in \mathbb{R}^n$ and $R > 0$. We will denote by $B(x,R) = \{y \in \mathbb{R}^n \mid \|x-y\| < R\}$ the open ball centered at x of radius R. The set $\overline{B(x,R)} = \{y \in \mathbb{R}^n \mid \|x-y\| \leq R\}$ is the closure $B(x,R)$.
- Let $U \subset \mathbb{R}^n$ be an open set. We will denote by $\mathcal{C}^k(U,\mathbb{R}^m)$ the set of k times continously differentiable functions $f\colon U \to \mathbb{R}^m$. If the codomain is clear from the context, we will use $\mathcal{C}^k(U)$. The set of continuous functions $f\colon U \to \mathbb{R}^m$ will be denoted $\mathcal{C}(U,\mathbb{R}^m)$ or $\mathcal{C}(U)$ when the codomain is clear.
- The velocity of a fluid will be the vector field $\mathbf{v} = \mathbf{v}(x,t)$. When working in $\mathbb{R}^2$ it will be written as $\mathbf{v} = u\mathbf{i} + v\mathbf{j}$.

Hereinafter, if $\mathcal{V} \subset \mathbb{R}^n$ is a control volume, we will assume it satisfies the following:

(i) $\mathcal{V}$ is an open set of $\mathbb{R}^n$, i.e. for all $x \in \mathcal{V}$ there exists $R > 0$ such that $B(x,R) \subset \mathcal{V}$.

(ii) $\mathcal{V}$ is bounded, that is to say, there exist $x_0 \in \mathbb{R}^n$ and $R > 0$ such that $\mathcal{V} \subset B(x_0,R)$.

(iii) $\mathcal{V}$ is a $\mathcal{C}^1$–domain. This implies that for every point $x \in \partial\mathcal{V}$ there exists a system of coordinates $(y_1,\ldots,y_{n-1},y_n) \equiv (\mathbf{y}',y_n)$ with origin at x, a ball $B(x,R)$ and a function φ defined in an open subset $\mathcal{N} \subset \mathbb{R}^{n-1}$ containing $\mathbf{y}' = \mathbf{0}'$, such that [1]:
   (a) $\varphi(\mathbf{0}') = 0$ and $\varphi \in \mathcal{C}^1(\mathcal{N},\mathbb{R})$ (φ is a $\mathcal{C}^1$ function from $\mathcal{N}$ to $\mathbb{R}$),
   (b) $\partial\mathcal{V} \cap B(x,R) = \{(\mathbf{y}',y_n) \mid y_n = \varphi(\mathbf{y}'),\ \mathbf{y}' \in \mathcal{N}\}$,
   (c) $\mathcal{V} \cap B(x,R) = \{(\mathbf{y}',y_n) \mid y_n > \varphi(\mathbf{y}'),\ \mathbf{y}' \in \mathcal{N}\}$.

Condition (i) will be useful to cast an integral equation into a differential equation. Condition (ii) prevents the integral of a continuous function defined on $\overline{\mathcal{V}}$ from becoming infinite. Moreover, an unbounded control volume, that is to say, a subset of $\mathbb{R}^n$ that extends indefinitely, makes no physical sense. Condition (iii), which is more technical, will allow us to apply vector calculus theorems.

Finally, we shall assume that all physical magnitudes, such as the velocity field $\mathbf{v}$, the density ρ or the temperature T are differentiable functions on their domains of definition as many times as necessary.

1.2 Reynolds Transport Theorem

Before stating and proving Reynolds Transport Theorem, we tackle the simpler Leibniz integral rule. To gain some physical intuition on it, suppose we have a very thin tube along the x–axis containing a fluid in motion. In this context we may assume that the fluid only moves along the tube direction. Let $f = f(x,t)$ be a magnitude of the fluid, for instance, the velocity u, the temperature T or the

concentration of some chemical species Y. So as to study how this magnitude variates on a portion of fluid, we consider a control volume $U(t) = [a(t), b(t)]$ that depends upon time. This situation is picture in figure 1.1. The total ammount of magnitude f in the control volume at time t, which we will denote by $\mathcal{F}(t)$, is given by

$$\mathcal{F}(t) = \int_{a(t)}^{b(t)} f(x,t)\,\mathrm{d}x \tag{1.1}$$

and its rate of variation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\int_{a(t)}^{b(t)} f(x,t)\,\mathrm{d}x \tag{1.2}$$

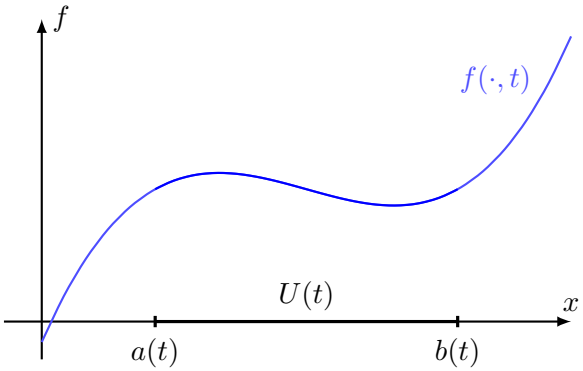


Figure 1.1. Control volume and magnitude f at time t.

Computing the derivative in equation (1.2) can be difficult depending on the case. Here is where Leibniz integral rule comes into play:

Theorem 1.1 (Leibniz integral rule)**.** Let $U \subset \mathbb{R}$ be a closed bounded interval and let $I = [t_1, t_2]$ be the time interval. Let $a, b\colon I \to U$ be differentiable functions with continuous derivative. Let $f\colon U \times I \to \mathbb{R}$, $(x,t) \mapsto f(x,t)$ be a differentiable function such that $\frac{\partial f}{\partial t}$ is also continuous. Then for all $t \in (t_1, t_2)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{a(t)}^{b(t)} f(x,t)\,\mathrm{d}x = \int_{a(t)}^{b(t)} \frac{\partial f}{\partial t}\,\mathrm{d}x + f(b(t),t)b'(t) - f(a(t),t)a'(t) \tag{1.3}$$

Proof. See [2]. □

Consider the more general case where we have a fluid in n–dimensional space $\mathbb{R}^n$ and magnitude $f = f(x,t)$ defined on a control volume $\mathcal{V}(t) \subset \mathbb{R}^n$. The total ammount of f on $\mathcal{V}$ at time t and its variation are given by similar formulas,

$$\mathcal{F}(t) = \int_{\mathcal{V}(t)} f(x,t)\,\mathrm{d}x\,, \quad \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}(t)} f(x,t)\,\mathrm{d}x \tag{1.4}$$

however now computing the derivative might be impracticable. In this case we have Reynolds Transport Theorem:

Theorem 1.2 (Reynolds Transport Theorem [3])**.** Let $U \subset \mathbb{R}^n$ be a compact set (i.e. U is closed and bounded) and let $\mathcal{V}(t)$ be a control volume depending on time such that $\mathcal{V} \subset U$ for all $t \in I = [0, T]$ with $T > 0$. Let $\mathcal{S}(t) = \partial\mathcal{V}(t)$ be the boundary of $\mathcal{V}(t)$ and let $F \in \mathcal{C}^1(U \times I, \mathbb{R})$ be a scalar field. Then for all $t \in I$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}(t)} F(x,t)\,\mathrm{d}x = \int_{\mathcal{V}(t)} \frac{\partial F}{\partial t}(x,t)\,\mathrm{d}x + \int_{\mathcal{S}(t)} F(x,t)\mathbf{b}\cdot\mathbf{n}\,\mathrm{d}S \tag{1.5}$$

where $\mathbf{b}\colon \mathcal{S}(t) \to \mathbb{R}^n$ is the local velocity of the control surface.

Proof. The moving control volume $\mathcal{V}(t)$ can be seen as the image of an initial region $\mathcal{V}(0)$ by a family of $\mathcal{C}^1$ maps $\xi\colon U \times I \subset \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, that is to say, $\mathcal{V}(t) = \xi(\mathcal{V}(0), t)$ for all $t \in I$. Furthermore, by fixing one time t, the mapping $\xi(\cdot, t)\colon \mathcal{V}(0) \to \mathcal{V}(t)$ can be assumed to be a diffeomorphism. Since F is continuous, we can apply the Change of Variables Theorem taking $x = \xi(x_0, t)$,

$$\int_{\mathcal{V}(t)} F(x,t)\,\mathrm{d}x = \int_{\mathcal{V}(0)} F(\xi(x_0,t),t)\left|\det\left(\frac{\partial \xi}{\partial x_0}(x_0,t)\right)\right|\mathrm{d}x_0$$

where the determinant of the jacobian matrix $\det\left(\frac{\partial \xi}{\partial x_0}(x_0,t)\right)$ can be assumed to be positive for small enough T, hence the absolute value is dropped. Applying differentiation under the integral sign (Theorem A.8) with respect to t yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}(t)} F(x,t)\,\mathrm{d}x &= \int_{\mathcal{V}(0)} \frac{\partial}{\partial t}\left\{F(\xi(x_0,t),t)\ \det\left(\frac{\partial \xi}{\partial x_0}(x_0,t)\right)\right\}\mathrm{d}x_0 \\ &= \int_{\mathcal{V}(0)} \frac{\partial}{\partial t}\left\{F(\xi(x_0,t),t)\right\}\det\left(\frac{\partial \xi}{\partial x_0}(x_0,t)\right)\mathrm{d}x_0 + \int_{\mathcal{V}(0)} F(\xi(x_0,t),t)\,\frac{\partial}{\partial t}\left\{\det\left(\frac{\partial \xi}{\partial x_0}(x_0,t)\right)\right\}\mathrm{d}x_0\end{aligned}$$

On the one hand,

$$\frac{\partial}{\partial t}\left\{F(\xi(x_0,t),t)\right\}\det\left(\frac{\partial \xi}{\partial x_0}(x_0,t)\right) = \left\{\frac{\partial F}{\partial t}(\xi(x_0,t),t) + \boldsymbol{\nabla} F(\xi(x_0,t),t)\cdot \xi_t(x_0,t)\right\}\det\left(\frac{\partial \xi}{\partial x_0}(x_0,t)\right)$$

where $\xi_t = \frac{\partial \xi}{\partial t}$. On the other hand, using matrix calculus,

$$F(\xi(x_0,t),t)\,\frac{\partial}{\partial t}\left\{\det\left(\frac{\partial \xi}{\partial x_0}(x_0,t)\right)\right\} = F(\xi(x_0,t),t)\det\left(\frac{\partial \xi}{\partial x_0}(x_0,t)\right)\boldsymbol{\nabla}\cdot\xi_t(x_0,t)$$

Thereby the integral is written as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}(t)} F(x,t)\,\mathrm{d}x &= \int_{\mathcal{V}(0)}\left\{\frac{\partial F}{\partial t} + \boldsymbol{\nabla}F\cdot\xi_t + F\boldsymbol{\nabla}\cdot\xi_t\right\}\det\left(\frac{\partial \xi}{\partial x_0}\right)\mathrm{d}x_0 \\ &= \int_{\mathcal{V}(0)}\left\{\frac{\partial F}{\partial t} + \boldsymbol{\nabla}\cdot(F\xi_t)\right\}\det\left(\frac{\partial \xi}{\partial x_0}\right)\mathrm{d}x_0\end{aligned}$$

So as to obtain an integral over $\mathcal{V}(t)$, the previous change of variables is reverted, that is, $x_0 = \xi^{-1}(x,t)$. In order not to complicate notation, let $\mathbf{b}(x,t) = \xi_t(\xi^{-1}(x,t),t)$, then

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}(t)} F(x,t)\,\mathrm{d}x = \int_{\mathcal{V}(t)}\left\{\frac{\partial F}{\partial t} + \boldsymbol{\nabla}\cdot(F\mathbf{b})\right\}(x,t)\,\mathrm{d}x$$

For a fixed $x_0 \in \mathcal{V}(0)$, $\xi(x_0,\cdot)$ is a function of time giving how x_0 moves, hence $\xi_t(x_0,t)$ is the instantaneous velocity of x_0. To end, an application of divergence theorem yields the final formula:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}(t)} F(x,t)\,\mathrm{d}x = \int_{\mathcal{V}(t)} \frac{\partial F}{\partial t}(x,t)\,\mathrm{d}x + \int_{\mathcal{S}(t)} F(x,t)\mathbf{b}\cdot\mathbf{n}\,\mathrm{d}S$$

□

1.3 Continuity equation

For the purposes of this project, where no nuclear nor relativistic effects are considered, mass is a property preserved over time. Let $\mathcal{V} \subset \mathbb{R}^n$ be a control volume, which may depend on time, and let

$\rho = \rho(x,t)$ be the mass density defined over $\mathcal{V}$ for each time $t \in I$. The mass enclosed by $\mathcal{V}$ at time t is

$$m(t) = \int_{\mathcal{V}(t)} \rho(x,t)\,\mathrm{d}x = \int_{\mathcal{V}(t)} \rho\,\mathrm{d}x \tag{1.6}$$

and as a result of the mass conservation principle

$$\frac{\mathrm{d}}{\mathrm{d}t} m(t) = \frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathcal{V}(t)} \rho(x,t)\,\mathrm{d}x = 0 \tag{1.7}$$

Now applying Reynolds Transport Theorem to (1.7) setting $\mathbf{b} = \mathbf{v}$,

$$\int_{\mathcal{V}(t)} \frac{\partial \rho}{\partial t}\,\mathrm{d}x + \int_{\mathcal{S}(t)} \rho\mathbf{v}\cdot\mathbf{n}\,\mathrm{d}S = 0 \tag{1.8}$$

We apply the divergence theorem on the surface integral to transform it into a volume integral,

$$\int_{\mathcal{V}(t)} \frac{\partial \rho}{\partial t}\,\mathrm{d}x + \int_{\mathcal{V}(t)} \boldsymbol{\nabla}\cdot(\rho\mathbf{v})\,\mathrm{d}x = \int_{\mathcal{V}(t)} \left\{\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\mathbf{v})\right\}\mathrm{d}x = 0 \tag{1.9}$$

We claim that the integrand in equation (1.9) vanishes at every point in space and time. Indeed, assume there exists a time t_0 and a point $x_0 \in \mathcal{V}(t_0)$ such that

$$\left\{\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\mathbf{v})\right\}\Big|_{(x_0,t_0)} > 0 \tag{1.10}$$

Recall that we assumed the physical magnitudes are differentiable functions as many times as necessary. In particular, by fixing $t = t_0$, $(\partial_t\rho + \boldsymbol{\nabla}\cdot(\rho\mathbf{v}))(\cdot,t_0)$ is a continuous function of x. Since $\mathcal{V}(t_0)$ is open, there exists $\tilde{\delta} > 0$ such that $B(x_0,\tilde{\delta}) \subset \mathcal{V}(t_0)$. By continuity we can take $\delta > 0$, with $\delta < \tilde{\delta}$ such that for all $y \in B(x_0,\delta) \subset \mathcal{V}(t_0)$,

$$\left\{\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\mathbf{v})\right\}\Big|_{(y,t_0)} > 0 \tag{1.11}$$

Hence integrating on $B(x_0,\delta)$ yields

$$\int_{B(x_0,\delta)} \left\{\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\mathbf{v})\right\}\mathrm{d}x > 0 \tag{1.12}$$

a contradiction as it should be zero according to equation (1.9). The same contradiction is reached if we assume the existence of a point x_0 and a time t_0 where

$$\left\{\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\mathbf{v})\right\}\Big|_{(x_0,t_0)} < 0 \tag{1.13}$$

thereby proving our claim. Because this is true for each $x_0 \in \mathcal{V}(t_0)$ and $t_0 \in I$ is arbitrary, we obtain the continuity equation

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\mathbf{v}) = 0 \tag{1.14}$$

1.4 General convection–diffusion equation

Let $\mathcal{V} \subset \mathbb{R}^n$ be a control volume which may depend on time $t \in I \subset \mathbb{R}$ and let $\phi\colon \mathbb{R}^n \times I \to \mathbb{R}$, $(x,t) \mapsto \phi(x,t)$ be a magnitude of the fluid (such as the concentration of some chemical substance) per unit of mass. Then the total ammount of ϕ in $\mathcal{V}(t)$ is

$$\Phi(t) = \int_{\mathcal{V}(t)} \rho(x,t)\phi(x,t)\,\mathrm{d}x \tag{1.15}$$

and its variation over time is

$$\dot{\Phi}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\Phi(t) = \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}(t)} \rho(x,t)\phi(x,t)\,\mathrm{d}x \tag{1.16}$$

The variation of Φ is a consequence of two contributions: the flux of ϕ through the control surface $\mathcal{S}(t)$ and the generation/elimination of ϕ in $\mathcal{V}(t)$ due to source terms. Let $\mathbf{f}\colon \mathbb{R}\times\mathcal{S}\times I \to \mathbb{R}^n$, $(\phi,x,t)\mapsto \mathbf{f}(\phi,x,t)$ be the vector field which gives the flux of ϕ through $\mathcal{S}$. Then the total ammount of ϕ flowing through $\mathcal{S}(t)$ is given by

$$\mathscr{F}(t) = \int_{\mathcal{S}(t)} \mathbf{f}(\phi,x,t)\cdot\mathbf{n}\,\mathrm{d}S \tag{1.17}$$

In order to find out which sign has $\mathscr{F}(t)$, we may assume for a moment that there are no source terms. If $\mathscr{F}(t)>0$, then ϕ is exiting $\mathcal{V}(t)$ and, as a result, $\dot{\Phi}(t)<0$. Conversely, if $\mathscr{F}(t)<0$, then $\dot{\Phi}(t)>0$, therefore $\dot{\Phi}(t)$ and $F(t)$ have opposite signs. Now, let $\dot{s}_\phi\colon \to\mathbb{R}$ be te source term, which provides the ammount of ϕ generated/eliminated in $\mathcal{V}(t)$ per unit of time. Then the total ammount of ϕ generated/eliminated in $\mathcal{V}(t)$ is

$$\mathscr{S}(t) = \int_{\mathcal{V}(t)} \dot{s}_\phi(\phi,x,t)\,\mathrm{d}x \tag{1.18}$$

Assume that there is no flux of ϕ through $\mathcal{S}(t)$, that is to say, $\mathscr{F}(t)=0$. If ϕ is generated in $\mathcal{V}(t)$, then $\mathscr{S}(t)>0$, which implies $\dot{\phi}(t)>0$; whereas if $\mathscr{S}(t)<0$ then $\dot{\phi}(t)>0$, thus $\dot{\phi}(t)$ and $\mathscr{S}(t)$ have the same sign. Introducing these terms in (1.16) leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}(t)} \rho(x,t)\phi(x,t)\,\mathrm{d}x = -\int_{\mathcal{S}(t)} \mathbf{f}(\phi,x,t)\cdot\mathbf{n}\,\mathrm{d}S + \int_{\mathcal{V}(t)} \dot{s}_\phi(\phi,x,t)\,\mathrm{d}x \tag{1.19}$$

Hereinafter we shall become less formal by omitting on which variables depends each function. In order to relate the flux $\mathbf{f}$ and ϕ, we need to apply some constitutive law. Fourier's law for heat conduction and Fick's law for concentration state that $\mathbf{f}$ depends linearly on the gradient of ϕ with respect to the spatial variables, that is,

$$\mathbf{f} = -\Gamma_\phi \boldsymbol{\nabla}_x\phi = -\Gamma_\phi \begin{pmatrix} \dfrac{\partial\phi}{\partial x_1} & \cdots & \dfrac{\partial\phi}{\partial x_n}\end{pmatrix}^T \tag{1.20}$$

where Γ_ϕ is known as the diffusion coefficient. So as not to complicate the notation, we will write $\boldsymbol{\nabla}\phi$ in place of $\boldsymbol{\nabla}_x\phi$. Recall that $\boldsymbol{\nabla}\phi\in\mathbb{R}^n$ gives the direction of maximum growth of $\phi(\cdot,t)$ (the time is fixed because the gradient is computed with respect to x). The minus sign in (1.20) is the consequence of heat (concentration of a chemical) flowing from regions of higher to lower temperature (concentration) regions. With this in mind, equation (1.19) is rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}(t)} \rho\phi\,\mathrm{d}x = \int_{\mathcal{S}(t)} \Gamma_\phi\boldsymbol{\nabla}\phi\cdot\mathbf{n}\,\mathrm{d}S + \int_{\mathcal{V}(t)} \dot{s}_\phi\,\mathrm{d}x \tag{1.21}$$

and applying Reynolds Transport Theorem on the left–hand side of (1.21) with $\mathbf{b}=\mathbf{v}$,

$$\int_{\mathcal{V}(t)} \frac{\partial(\rho\phi)}{\partial t}\,\mathrm{d}x + \int_{\mathcal{S}(t)} \rho\phi\mathbf{v}\cdot\mathbf{n}\,\mathrm{d}S = \int_{\mathcal{S}(t)} \Gamma_\phi\boldsymbol{\nabla}\phi\cdot\mathbf{n}\,\mathrm{d}S + \int_{\mathcal{V}(t)} \dot{s}_\phi\,\mathrm{d}x \tag{1.22}$$

To turn surface integrals into volume integrals we apply divergence theorem,

$$\int_{\mathcal{V}(t)} \frac{\partial(\rho\phi)}{\partial t}\,\mathrm{d}x + \int_{\mathcal{V}(t)} \boldsymbol{\nabla}\cdot(\rho\phi\mathbf{v})\,\mathrm{d}x = \int_{\mathcal{V}(t)} \boldsymbol{\nabla}\cdot(\Gamma_\phi\boldsymbol{\nabla}\phi)\,\mathrm{d}x + \int_{\mathcal{V}(t)} \dot{s}_\phi\,\mathrm{d}x \tag{1.23}$$

Proceeding in a similar way to the continuity equation, we assume the existence of a time t_0 and a point x_0 where

$$\left\{ \frac{\partial(\rho\phi)}{\partial t}\,\mathrm{d}x + \boldsymbol{\nabla}\cdot(\rho\phi\mathbf{v})\,\mathrm{d}x - \boldsymbol{\nabla}\cdot(\Gamma_\phi\boldsymbol{\nabla}\phi)\,\mathrm{d}x - \dot{s}_\phi\,\mathrm{d}x \right\}\Big|_{(x_0,t_0)} \neq 0 \tag{1.24}$$

and we reach a contradiction, thereby obtaining the general convection diffusion equation:

$$\frac{\partial(\rho\phi)}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\phi\mathbf{v}) = \boldsymbol{\nabla}\cdot(\Gamma_\phi\boldsymbol{\nabla}\phi) + \dot{s}_\phi \tag{1.25}$$

The left–hand side of (1.25) can be expanded to find

$$\phi\left\{\frac{\partial\rho}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\mathbf{v})\right\} + \rho\frac{\partial\phi}{\partial t} + \rho\mathbf{v}\cdot\boldsymbol{\nabla}\phi = \boldsymbol{\nabla}\cdot(\Gamma_\phi\boldsymbol{\nabla}\phi) + \dot{s}_\phi \tag{1.26}$$

Since the term between keys is the continuity equation, (1.26) is simplified to

$$\rho\frac{\partial\phi}{\partial t} + \rho\mathbf{v}\cdot\boldsymbol{\nabla}\phi = \boldsymbol{\nabla}\cdot(\Gamma_\phi\boldsymbol{\nabla}\phi) + \dot{s}_\phi \tag{1.27}$$

Equations (1.25) and (1.27) are two equivalent forms of the same equation, each having its applications and benefits.

By taking ϕ to be the temperature T of the fluid or the concentration of the k–th chemical substance Y_k in the fluid, one obtains the energy conservation equation (1.28) and the k–species equation (1.29) [4]:

$$\frac{\partial(\rho T)}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\mathbf{v}T) = \boldsymbol{\nabla}\cdot\left(\frac{\lambda}{c_v}\boldsymbol{\nabla}T\right) + \left\{\frac{\tau\circ\boldsymbol{\nabla}\mathbf{v} - \boldsymbol{\nabla}\cdot\dot{\mathbf{q}}^R - p\boldsymbol{\nabla}\cdot\mathbf{v}}{c_v}\right\} \tag{1.28}$$

$$\frac{\partial(\rho Y_k)}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\mathbf{v}Y_k) = \boldsymbol{\nabla}\cdot(\rho D_{km}\boldsymbol{\nabla}Y_k) + \{\dot{\omega}_k\} \tag{1.29}$$

If ϕ is not a scalar magnitude but a vector magnitude, i.e. $\phi \equiv (\phi_1,\ldots,\phi_n)\colon \mathbb{R}^n\times I \to \mathbb{R}^n$, $(x,t)\mapsto\phi(x,t)$, the same process applied on each component function ϕ_i leads to n equations similar to (1.25) that can be gathered in the following vector equation

$$\frac{\partial(\rho\phi)}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\mathbf{v}\otimes\phi) = \boldsymbol{\nabla}\cdot(\mu\boldsymbol{\nabla}\phi) + \dot{s}_\phi \tag{1.30}$$

where $\mathbf{v}\otimes\phi$ is the exterior product:

$$\mathbf{v}\otimes\phi = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}\begin{pmatrix} \phi_1 & \cdots & \phi_n \end{pmatrix} = \begin{pmatrix} v_1\phi_1 & \cdots & v_1\phi_n \\ \vdots & \ddots & \vdots \\ v_n\phi_1 & \cdots & v_n\phi_n \end{pmatrix} \tag{1.31}$$

The previous product can also be regarded as the tensor product of two 1–covariant tensors which yields a 2–covariant tensor. Notice that, in general, this product is not commutative, that is to say, $\mathbf{v}\otimes\phi \neq \phi\otimes\mathbf{v}$.

By taking ϕ to be the velocity $\mathbf{v}$ of the fluid, the momentum conservation equation is obtained [4]:

$$\frac{\partial(\rho\mathbf{v})}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\mathbf{v}\otimes\mathbf{v}) = \boldsymbol{\nabla}\cdot(\mu\boldsymbol{\nabla}\mathbf{v}) + \{\boldsymbol{\nabla}\cdot(\tau - \mu\boldsymbol{\nabla}\mathbf{v}) - \boldsymbol{\nabla}p + \rho\mathbf{g}\} \tag{1.32}$$

2 Two important Partial Differential Equations

In this section we introduce two important partial differential equations, namely, the transport equation and the diffusion equation. We shall not deduce them nor work out the solution to an initial value problem, but rather gain physical intuition about the problem they model and the solution, as well as discuss their relation to the convection–diffusion equations. The references for this section are [1] and [5].

We will start by introducing some notation:

- A vector $\alpha = (\alpha_1, \dots, \alpha_n)$ such that $\alpha_1, \dots, \alpha_n \geq 0$ is called a multiindex of order $|\alpha| = \alpha_1 + \cdots + \alpha_n$.
- Given a multiindex α and a sufficiently differentiable function u, we define

$$D^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}} = \partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n} u \tag{2.1}$$

- If $k \in \mathbb{Z}_{\geq 0}$, we denote by

$$D^k u(x) = \{D^\alpha u(x) \mid |\alpha| = k\} \tag{2.2}$$

the set of all partial derivatives of order k. For $k = 1$ we have the gradient vector $Du = \boldsymbol{\nabla} u = (\partial_{x_1} u, \dots, \partial_{x_n} u)^\top$ and for $k = 2$ we have the Hessian matrix,

$$D^2 u = \begin{pmatrix} \partial_{x_1}^2 u & \cdots & \partial_{x_1}\partial_{x_n} u \\ \vdots & \ddots & \vdots \\ \partial_{x_n}\partial_{x_1} u & \cdots & \partial_{x_n}^2 u \end{pmatrix} \tag{2.3}$$

- The symbol Δ denotes Laplace's operator, i.e. $\Delta u = \sum_{i=1}^n \partial_{x_i}^2 u$.

Informally, a partial differential equation (PDE) is an equation involving an unknown multivariable function and its partial derivatives. Hereinafter, let U be an open subset of $\mathbb{R}^n$,

2.1 The transport equation

The transport equation with constant coefficients is the following first–order linear partial differential equation

$$\frac{\partial u}{\partial t} + b \cdot Du = 0 \quad \text{in } \mathbb{R}^n \times (0, \infty) \tag{2.4}$$

where $b = (b_1, \dots, b_n)^\top \in \mathbb{R}^n$ is a fixed vector and $u \colon \mathbb{R}^n \times [0, \infty) \to \mathbb{R}$ is the unknown function, $u = u(x_1, \dots, x_n, t) = u(x, t)$. Du is the gradient vector with respect to the spatial variables, $Du = (\partial_{x_1} u, \dots, \partial_{x_n} u)$.

The initial value problem for the transport equation is

$$\begin{cases} \dfrac{\partial u}{\partial t} + b \cdot Du = 0 & \text{in } \mathbb{R}^n \times (0, \infty) \\ u = g & \text{on} \mathbb{R}^n \times \{t = 0\} \end{cases} \tag{2.5}$$

where $g \colon \mathbb{R}^n \to \mathbb{R}$ is a given function.

2.2 The diffusion equation

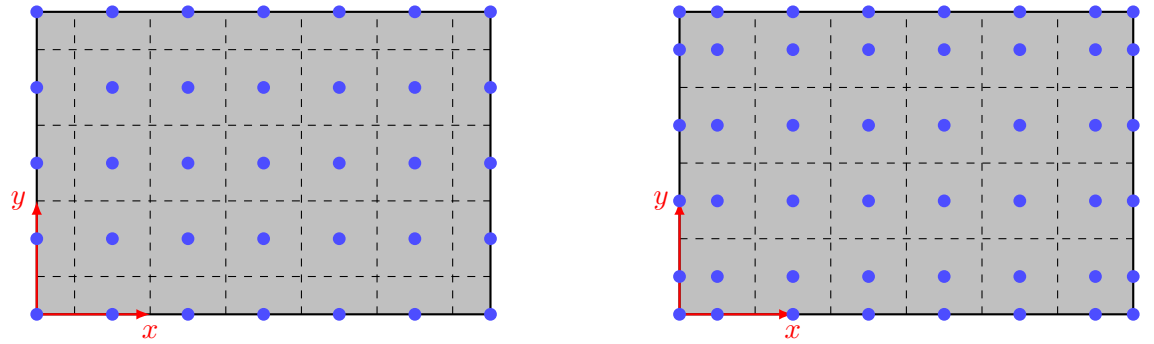
2.3 The diffusion equation

3 Numerical study of the convection–diffusion equations

3.1 Spatial and time discretization

The type of problems we will addressed in this project occur in a bounded rectangular domain $\Omega \subset \mathbb{R}^2$, that is to say, there exist non–degenerate intervals $[0, L]$ and $[0, W]$ such that $\Omega = [0, L] \times [0, W]$. In order to solve the problem numerically we shall follow a control–volume formulation. This methodology discretizes the domain into nonoverlapping control volumes along with a grid of points named discretization nodes. The resulting discretized domain is named mesh or numerical grid [6].

There exist several types of grids according to the shape of control volumes and the ammount of subdivisions the domain has been partitioned into, namely, a structured (regular) grid, a block–structured grid and an unstructured grid [7]. However, henceforth we will only consider structured regular grids. This formulation allows for two manners to discretize the domain, namely, cell–centered and node–centered discretizations. The former places discretization nodes over the domain and generates a control–volume centered on each node. The latter first generates the control–volumes, next places a node at the center of each one and finally sets nodes at the border if necessary.



(a) Cell–centered uniform discretization. **(b)** Node–centered uniform discretization.

Figure 3.1. Comparison of the cell–centered and the node–centered uniform discretizations.

As it can be noticed when uniform discretizations are used, the node–centered discretization approach offers higher resolution near the boundary of the domain. Notwithstanding, it also generates singular nodes located at the corners which need a special treatment, whilst the cell–centered does not. Furthermore, we can distinguish between uniform discretizations, where distances between adjacent internal nodes are constant along the domain, and non–uniform discretizations, meaning the opposite.

Later we will deal with the discretization of the convection–diffusion equations in a cell–centered discretized domain. In order to enumerate the nodes, we will start from the lower left corner of Ω, where node $(0, 0)$ is located. We will use the notation (i, j) to refer the i–th node in x–coordinate and j–th node in y–coordinate. Given an arbitrary node (i, j) that we denote by P, its neighbour nodes are the west node $(i - 1, j)$, the east node $(i + 1, j)$, the south node $(i, j - 1)$ and the north node $(i, j + 1)$. This scheme is pictured in figure 3.2. The calligraphic letter $\mathcal{V}$ will be used to denote a control volume. For instance, $\mathcal{V}_P$ is the control volume associated to node P. The volume of $\mathcal{V}_P$ is V_P. The notation $\mathcal{S}_{Pi}$ will denote the interface between the control volumes $\mathcal{V}_P$ and $\mathcal{V}_I$. As an example, $\mathcal{S}_{Pw}$ is the surface between the control volumes $\mathcal{V}_P$ and $\mathcal{V}_W$. The distance between the control volumes associated to nodes A and B is $d_{AB} = \|(x_A - x_B, y_A - y_B)\|$. The distance between the node P and one of its control surfaces i is given by d_{Pi}, for example, d_{Pw}.

In regards to time, the problems we consider last for finite time. Therefore the time interval is $I = [0, T] \subset \mathbb{R}$ with $T > 0$ finite. The discretization of I is simply a partition of it, that is to say, a

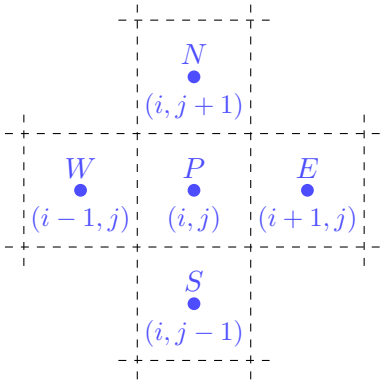


Figure 3.2. Central node P and neighbour nodes.

finite set of points $P(I) = \{t_0 = 0, t_1, \ldots, t_{m-1}, t_m = T\}$ with $t_{i+1} > t_i$ for all $0 \leq i < m$. The time discretization is said to uniform whenever there exists $\Delta t > 0$ such that $t_{i+1} - t_i = \Delta t$ for all i, and non–uniform otherwise. In the case of a uniform time discretization, the number Δt is know as time step. We shall only consider uniform time discretizations, nevertheless non–uniform discretizations might be convenient in problems combining fast and low transient processes.

3.2 Discretization of the continuity equation

As we have previously seen, the continuity equation in differential form is

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot (\rho \mathbf{v}) = 0 \quad (x, t) \in \Omega \times I \tag{3.1}$$

Since the above relation is true in $\Omega \times I$, fixing one time $t \in I$ and integrating over a control volume $\mathcal{V}_P \subset \Omega$ gives

$$\int_{\mathcal{V}_P} \frac{\partial \rho}{\partial t} \, \mathrm{d}x + \int_{\mathcal{V}_P} \boldsymbol{\nabla} \cdot (\rho \mathbf{v}) \, \mathrm{d}x = 0 \tag{3.2}$$

Let $\mathcal{S}_P = \partial \mathcal{V}_P$ be the control surface, i.e. the boundary of the control volume. Then applying the divergence theorem on the second term of equation (3.2),

$$\int_{\mathcal{V}_P} \frac{\partial \rho}{\partial t} \, \mathrm{d}x + \int_{\mathcal{S}_P} \rho \mathbf{v} \cdot \mathbf{n} \, \mathrm{d}S = 0 \tag{3.3}$$

So as to simplify the first term of (3.3), we define the average density of the control volume as

$$\overline{\rho}_P = \frac{1}{V_P} \int_{\mathcal{V}_P} \rho \, \mathrm{d}x \tag{3.4}$$

Introducing this relation in equation (3.3),

$$\frac{\mathrm{d}\overline{\rho}_P}{\mathrm{d}t} V_P + \int_{\mathcal{S}_P} \rho \mathbf{v} \cdot \mathbf{n} \, \mathrm{d}S = 0 \tag{3.5}$$

The mass flow term can be further simplified because we know the geometry of the boundary of $\mathcal{V}_P$. Since the control surface is $\mathcal{S}_P = \mathcal{S}_{Pe} \cup \mathcal{S}_{Pw} \cup \mathcal{S}_{Pn} \cup \mathcal{S}_{Ps}$, we can rewrite the mass flow term as

$$\begin{aligned} \int_{\mathcal{S}_P} \rho \mathbf{v} \cdot \mathbf{n} \, \mathrm{d}S &= \underbrace{\int_{\mathcal{S}_{Pe}} \rho \mathbf{v} \cdot \mathbf{n} \, \mathrm{d}S}_{\dot{m}_e} + \underbrace{\int_{\mathcal{S}_{Pw}} \rho \mathbf{v} \cdot \mathbf{n} \, \mathrm{d}S}_{-\dot{m}_w} + \underbrace{\int_{\mathcal{S}_{Pn}} \rho \mathbf{v} \cdot \mathbf{n} \, \mathrm{d}S}_{\dot{m}_n} + \underbrace{\int_{\mathcal{S}_{Ps}} \rho \mathbf{v} \cdot \mathbf{n} \, \mathrm{d}S}_{-\dot{m}_s} \\ &= \dot{m}_e - \dot{m}_w + \dot{m}_n - \dot{m}_s \end{aligned} \tag{3.6}$$

Since evaluating each integral may be computationally expensive or impossible, the following approach is followed. Given a face f, the normal outer vector is constant on $\mathcal{S}_{Pf}$. Indeed, if $\mathbf{n}_f$ denotes the normal outer vector to face f, then $\mathbf{n}_e = \mathbf{i}$, $\mathbf{n}_w = -\mathbf{i}$, $\mathbf{n}_n = \mathbf{j}$ and $\mathbf{n}_s = -\mathbf{j}$. Since $\mathbf{v} = u\mathbf{i} + v\mathbf{j}$, the dot products are $\mathbf{v}\cdot\mathbf{n}_e = u$, $\mathbf{v}\cdot\mathbf{n}_w = -u$ and so on. Moreover, the integrand $(\rho\mathbf{v}\cdot\mathbf{n})_f$ can be approximated by the value each term takes at the face center, i.e.

$$(\rho\mathbf{v}\cdot\mathbf{n})_f \approx \rho_f(\mathbf{v}\cdot\mathbf{n})_f \tag{3.7}$$

Therefore the integral over $\mathcal{S}_{Pe}$ on equation (3.6) is simplified as follows:

$$\int_{\mathcal{S}_{Pe}} \rho\mathbf{v}\cdot\mathbf{n}\,\mathrm{d}S \approx \int_{\mathcal{S}_{Pe}} (\rho\mathbf{v}\cdot\mathbf{n})_e\,\mathrm{d}S \approx \int_{\mathcal{S}_{Pe}} \rho_e(\mathbf{v}\cdot\mathbf{n})_e)\,\mathrm{d}S = \int_{\mathcal{S}_{Pe}} \rho_e u_e\,\mathrm{d}S = \rho_e u_e S_{Pe} =: \dot{m}_e \tag{3.8}$$

The same simplifications are applied to the other integrals. Defining $\dot{m}_w$ and $\dot{m}_s$ as the negative integral makes the mass flow terms be positive in the positive coordinate direction. Introducing these in (3.5) yields

$$\frac{\mathrm{d}\overline{\rho}_P}{\mathrm{d}t}V_P + \dot{m}_e - \dot{m}_w + \dot{m}_n - \dot{m}_s = 0 \tag{3.9}$$

The average density of the control volume is roughly the density at the discretization node, i.e. $\overline{\rho}_P \approx \rho_P$. Integrating (3.9) over the time interval $[t^n, t^{n+1}]$ gives

$$V_P \int_{t^n}^{t^{n+1}} \frac{\mathrm{d}\rho_P}{\mathrm{d}t}\,\mathrm{d}t + \int_{t^n}^{t^{n+1}} (\dot{m}_e - \dot{m}_w + \dot{m}_n - \dot{m}_s)\,\mathrm{d}t = 0 \tag{3.10}$$

The first term of (3.10) has a straightforward simplification applying a corollary of the fundamental theorem of calculus. Regarding the second term, numerical integration is used,

$$\begin{aligned}(\rho_P^{n+1} - \rho_P^n)V_P + \beta(\dot{m}_e^{n+1} - \dot{m}_w^{n+1} + \dot{m}_n^{n+1} - \dot{m}_s^{n+1})(t^{n+1} - t^n) \\ + (1-\beta)(\dot{m}_e^n - \dot{m}_w^n + \dot{m}_n^n - \dot{m}_s^n)(t^{n+1} - t^n) = 0\end{aligned} \tag{3.11}$$

where $\beta \in \{0, \frac{1}{2}, 1\}$ depends on the chosen integration scheme. For the sake of simplicity, superindex $n+1$ shall be dropped and the time instant n will be denoted by the superindex 0. Since we assume a uniform time discretization with time step Δt, the resulting discretized continuity equation is

$$\frac{\rho_P - \rho_P^0}{\Delta t}V_P + \beta(\dot{m}_e - \dot{m}_w + \dot{m}_n - \dot{m}_s) + (1-\beta)(\dot{m}_e^0 - \dot{m}_w^0 + \dot{m}_n^0 - \dot{m}_s^0) = 0 \tag{3.12}$$

Finally, when an implicit scheme is selected for the time integration,

$$\frac{\rho_P - \rho_P^0}{\Delta t}V_P + \dot{m}_e - \dot{m}_w + \dot{m}_n - \dot{m}_s = 0 \tag{3.13}$$

If a 3D–mesh is being used, the contributions of top (T) and bottom (B) nodes must be considered. In this case, the control surface is the union $\mathcal{S}_P = \mathcal{S}_{Pe} \cup \mathcal{S}_{Pw} \cup \mathcal{S}_{Pn} \cup \mathcal{S}_{Ps} \cup \mathcal{S}_{Pt} \cup \mathcal{S}_{Pb}$, hence equation (3.13) incorporates two new terms

$$\frac{\rho_P - \rho_P^0}{\Delta t}V_P + \dot{m}_e - \dot{m}_w + \dot{m}_n - \dot{m}_s + \dot{m}_t - \dot{m}_b = 0 \tag{3.14}$$

3.3 Discretization of the general convection–diffusion equation

The generalized convection–diffusion for a real valued function $\phi\colon \Omega \times I \subset \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is

$$\frac{\partial(\rho\phi)}{\partial t} + \boldsymbol{\nabla}\cdot(\rho\mathbf{v}\phi) = \boldsymbol{\nabla}\cdot(\Gamma_\phi\boldsymbol{\nabla}\phi) + \dot{s}_\phi, \quad (x,t) \in \Omega \times I \tag{3.15}$$

whereas for a vector valued function $\phi = (\phi_1, \ldots, \phi_n)\colon \Omega \times I \subset \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ it is written as

$$\frac{\partial(\rho\phi)}{\partial t} + \boldsymbol{\nabla} \cdot (\rho \mathbf{v} \otimes \phi) = \boldsymbol{\nabla} \cdot (\Gamma_\phi \boldsymbol{\nabla}\phi) + \dot{s}_\phi, \quad (x,t) \in \Omega \times I \tag{3.16}$$

where $\otimes$ denotes the outer product of $\mathbf{v}\colon \Omega \times I \subset \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ and ϕ, which is a $n \times n$ matrix. Since the generalized convection–diffusion equation for a vector valued function actually comprises n equations, one for each component function, we will only study the discretization for a real valued function.

Integrating (3.15) over $\mathcal{V}_P \times [t^n, t^{n+1}] \subset \Omega \times I$ and using Fubini's theorem to swap the order of integration

$$\int_{t^n}^{t^{n+1}} \int_{\mathcal{V}_P} \frac{\partial(\rho\phi)}{\partial t} \,\mathrm{d}x\,\mathrm{d}t + \int_{t^n}^{t^{n+1}} \int_{\mathcal{V}_P} \boldsymbol{\nabla} \cdot (\rho \mathbf{v} \phi) \,\mathrm{d}x\,\mathrm{d}t = \\ = \int_{t^n}^{t^{n+1}} \int_{\mathcal{V}_P} \boldsymbol{\nabla} \cdot (\Gamma_\phi \boldsymbol{\nabla}\phi) \,\mathrm{d}x\,\mathrm{d}t + \int_{t^n}^{t^{n+1}} \int_{\mathcal{V}_P} \dot{s}_\phi \,\mathrm{d}x\,\mathrm{d}t \tag{3.17}$$

The simplification of the first term is analogous to that of the continuity equation. The average value of $\rho\phi$ on $\mathcal{V}_P$ at time t is defined by

$$(\rho\phi)_P = \frac{1}{V_P} \int_{\mathcal{V}_P} \rho\phi \,\mathrm{d}x \tag{3.18}$$

although the following approximation is needed:

$$(\rho\phi)_P \approx \rho_P \phi_P \tag{3.19}$$

Then the transient term is:

$$\int_{t^n}^{t^{n+1}} \int_{\mathcal{V}_P} \frac{\partial(\rho\phi)}{\partial t} \,\mathrm{d}x\,\mathrm{d}t = \int_{t^n}^{t^{n+1}} \frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathcal{V}_P} \rho\phi \,\mathrm{d}x\,\mathrm{d}t = \int_{t^n}^{t^{n+1}} \frac{\mathrm{d}(\rho\phi)_P}{\mathrm{d}t} V_P \,\mathrm{d}t \approx \left\{\rho_P \phi_P - \rho_P^0 \phi_P^0\right\} V_P \tag{3.20}$$

Divergence theorem must be applied to simplify the convective term,

$$\int_{t^n}^{t^{n+1}} \int_{\mathcal{V}_P} \boldsymbol{\nabla} \cdot (\rho \mathbf{v} \phi) \,\mathrm{d}x\,\mathrm{d}t = \int_{t^n}^{t^{n+1}} \int_{\mathcal{S}_P} \rho\phi \mathbf{v} \cdot \mathbf{n} \,\mathrm{d}S\,\mathrm{d}t = \int_{t^n}^{t^{n+1}} \sum_i \int_{\mathcal{S}_{Pi}} \rho\phi \mathbf{v} \cdot \mathbf{n} \,\mathrm{d}S\,\mathrm{d}t \tag{3.21}$$

The value that ϕ takes on $\mathcal{S}_{Pi}$ can be approximated by its value at a representative point, for instance, the point at face center, that is to say, $\phi \approx \phi_i$. Therefore,

$$\int_{t^n}^{t^{n+1}} \sum_i \int_{\mathcal{S}_i} \rho\phi \mathbf{v} \cdot \mathbf{n} \,\mathrm{d}S\,\mathrm{d}t \approx \int_{t^n}^{t^{n+1}} \sum_i \int_{\mathcal{S}_i} \rho\phi_i \mathbf{v} \cdot \mathbf{n} \,\mathrm{d}S\,\mathrm{d}t = \int_{t^n}^{t^{n+1}} \sum_i \dot{m}_i \phi_i \,\mathrm{d}t = \\ = \left\{\beta \sum_i \dot{m}_i \phi_i + (1-\beta) \sum_i \dot{m}_i^0 \phi_i^0\right\} \Delta t \tag{3.22}$$

In regards to the diffusion term,

$$\int_{t^n}^{t^{n+1}} \int_{\mathcal{V}_P} \boldsymbol{\nabla} \cdot (\Gamma_\phi \boldsymbol{\nabla}\phi) \,\mathrm{d}x\,\mathrm{d}t = \int_{t^n}^{t^{n+1}} \int_{\mathcal{S}_P} \Gamma_\phi \boldsymbol{\nabla}\phi \cdot \mathbf{n} \,\mathrm{d}S\,\mathrm{d}t = \int_{t^n}^{t^{n+1}} \sum_i \int_{\mathcal{S}_{Pi}} \Gamma_\phi \boldsymbol{\nabla}\phi \cdot \mathbf{n} \,\mathrm{d}S\,\mathrm{d}t \tag{3.23}$$

The outer normal vector to the face $\mathcal{S}_{Pi}$ is constant and points in the direction of some coordinate axis, hence the dot product $\boldsymbol{\nabla}\phi \cdot \mathbf{n}$ in the face $\mathcal{S}_{Pi}$ equals the partial derivative with respect to x_i

times ± 1, depending on the direction of $\mathbf{n}$. For east, north and top faces the sign is positive, whilst for west, south and bottom faces the sign is negative. Again, Γ_ϕ will be approximated by the value at the face center, and partial derivatives will be approximated by a finite centered difference. In order to simplify the notation, we shall drop the subindex ϕ in the diffusion coefficient Γ_ϕ and define the coefficients

$$D_f = \frac{\Gamma_f S_f}{d_{PF}} \tag{3.24}$$

$$D_f^0 = \frac{\Gamma_f^0 S_f}{d_{PF}} \tag{3.25}$$

where f and F refer to the face and to the node, respectively. For a 2D–mesh, equation (3.23) results in

$$\begin{aligned}
&\int_{t^n}^{t^{n+1}} \sum_i \int_{\mathcal{S}_{Pi}} \Gamma_\phi \boldsymbol{\nabla}\phi \cdot \mathbf{n}\, \mathrm{d}S\, \mathrm{d}t \approx \\
&\approx \int_{t^n}^{t^{n+1}} \Big\{ D_e(\phi_E - \phi_P) - D_w(\phi_P - \phi_W) + D_n(\phi_N - \phi_P) - D_s(\phi_P - \phi_S) \Big\}\, \mathrm{d}t \approx \\
&\approx \beta \left\{ D_e(\phi_E - \phi_P) - D_w(\phi_P - \phi_W) + D_n(\phi_N - \phi_P) - D_s(\phi_P - \phi_S) \right\} \Delta t + \\
&+ (1-\beta) \Big\{ D_e^0(\phi_E - \phi_P) - D_w^0(\phi_P - \phi_W) + D_n^0(\phi_N - \phi_P) - D_s^0(\phi_P - \phi_S) \Big\} \Delta t
\end{aligned} \tag{3.26}$$

In the case of a 3D–mesh, the contributions of top and bottom faces must be accounted for.

So as to discretize the source term, the mean value of the source function in $\mathcal{V}_P$ at time t is given by

$$\overline{\dot{s}}_\phi = \frac{1}{V_P} \int_{\mathcal{V}_P} \dot{s}_\phi \, \mathrm{d}x \tag{3.27}$$

If the value of s_ϕ is known, the relation $\overline{\dot{s}}_\phi = \dot{\overline{s}}_\phi$ is true. Indeed, applying differentiation under the integral sign (Theorem A.8)

$$\dot{\overline{s}}_\phi = \frac{\mathrm{d}}{\mathrm{d}t}\overline{s}_\phi = \frac{1}{V_P}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}_P} s_\phi \, \mathrm{d}x = \frac{1}{V_P}\int_{\mathcal{V}_P} \dot{s}_\phi \, \mathrm{d}x = \overline{\dot{s}}_\phi \tag{3.28}$$

In most cases, the dependence of $\dot{\overline{s}}_\phi$ on ϕ is complicated. Since the equations obtained until now are linear, the relation between the source term and the variable would ideally be linear. This linearity is imposed as follows

$$\dot{\overline{s}}_\phi = S_C^\phi + S_P^\phi \phi_P \tag{3.29}$$

where the values of S_C^ϕ and S_P^ϕ may vary with ϕ [6]. Making use of these relations, the source term integral is discretized as

$$\int_{t^n}^{t^{n+1}} \int_{\mathcal{V}_P} \dot{s}_\phi \, \mathrm{d}x \, \mathrm{d}t = \int_{t^n}^{t^{n+1}} \dot{\overline{s}}_{\phi P} V_P \Delta t = \left(S_C^\phi + S_P^\phi \phi_P \right) V_P \Delta t \tag{3.30}$$

As we shall discuss later, the term S_P^ϕ must be non–positive.

The discretization of the 2D generalized convection–diffusion equation is

$$\begin{aligned}
&\frac{\rho_P\phi_P - \rho_P^0\phi_P^0}{\Delta t}V_P + \\
&+ \beta\Big(\dot{m}_e\phi_e - \dot{m}_w\phi_w + \dot{m}_n\phi_n - \dot{m}_s\phi_s\Big) + (1-\beta)\left(\dot{m}_e^0\phi_e^0 - \dot{m}_w^0\phi_w^0 + \dot{m}_n^0\phi_n^0 - \dot{m}_s^0\phi_s^0\right) = \\
&= \beta\Big\{D_e(\phi_E - \phi_P) - D_w(\phi_P - \phi_W) + D_n(\phi_N - \phi_P) - D_s(\phi_P - \phi_S)\Big\} + \\
&+ (1-\beta)\Big\{D_e^0(\phi_E^0 - \phi_P^0) - D_w^0(\phi_P^0 - \phi_W^0) + D_n^0(\phi_N^0 - \phi_P^0) - D_s^0(\phi_P^0 - \phi_S^0)\Big\} + \\
&+ \left(S_C^\phi + S_P^\phi\phi_P\right)V_P
\end{aligned} \tag{3.31}$$

In the case of a implicit integration scheme, i.e. $\beta = 1$, equation (3.31) is simplified to:

$$\begin{aligned}
&\frac{\rho_P\phi_P - \rho_P^0\phi_P^0}{\Delta t}V_P + \dot{m}_e\phi_e - \dot{m}_w\phi_w + \dot{m}_n\phi_n - \dot{m}_s\phi_s = \\
&\quad = D_e(\phi_E - \phi_P) - D_w(\phi_P - \phi_W) + D_n(\phi_N - \phi_P) - D_s(\phi_P - \phi_S) + (S_C^\phi + S_P^\phi\phi_P)V_P
\end{aligned} \tag{3.32}$$

An equivalent and more useful form of the discretization equation can be found by multiplying (3.13) by ϕ_P and subtracting it from (3.32), which results in

$$\begin{aligned}
&\rho_P^0\frac{\phi_P - \phi_P^0}{\Delta t}V_P + \dot{m}_e(\phi_e - \phi_P) - \dot{m}_w(\phi_w - \phi_P) + \dot{m}_n(\phi_n - \phi_P) - \dot{m}_s(\phi_s - \phi_P) = \\
&\quad = D_e(\phi_E - \phi_P) - D_w(\phi_P - \phi_W) + D_n(\phi_N - \phi_P) - D_s(\phi_P - \phi_S) + (S_C^\phi + S_P^\phi\phi_P)V_P
\end{aligned} \tag{3.33}$$

The 3D analog of (3.33) includes the top and bottom faces contributions:

$$\begin{aligned}
\rho_P^0\frac{\phi_P - \phi_P^0}{\Delta t}V_P &+ \dot{m}_e(\phi_e - \phi_P) - \dot{m}_w(\phi_w - \phi_P) + \dot{m}_n(\phi_n - \phi_P) \\
&- \dot{m}_s(\phi_s - \phi_P) + \dot{m}_t(\phi_t - \phi_P) - \dot{m}_b(\phi_b - \phi_P) \\
&= D_e(\phi_E - \phi_P) - D_w(\phi_P - \phi_W) + D_n(\phi_N - \phi_P) \\
&- D_s(\phi_P - \phi_S) + D_t(\phi_T - \phi_P) - D_b(\phi_P - \phi_B) + (S_C^\phi + S_P^\phi\phi_P)V_P
\end{aligned} \tag{3.34}$$

3.4 Evaluation of the convective terms

The discretized version of the generalized convection–diffusion equation requires the values of ϕ at points different from the nodes. In this subsection we study several methods to compute the value of ϕ at faces. We will assume that ρ and Γ are known at the nodal points. For the sake of simplicity, we will evaluate the convective term at east face, although the generalization to the remaining faces is straightforward. The references for this subsection are [6] and [8].

3.4.1 Upwind–Difference Scheme (UDS)

Incompressible flows and gases at low Mach number are more influenced by upstream conditions than downstream conditions. Let $(\mathbf{v}\cdot\mathbf{n})_e$ denote the value of the dot product $\mathbf{v}\cdot\mathbf{n}$ at east face $\mathcal{S}_{Pe}$. If $(\mathbf{v}\cdot\mathbf{n})_e > 0$, fluid flows from node P to node E, hence P is the upstream node and E is the downstream node. Conversely, if $(\mathbf{v}\cdot\mathbf{n})_e < 0$, nodes interchange their roles as fluid flows from node E to node P. Whenever $(\mathbf{v}\cdot\mathbf{n})_e = 0$, it implies that $\mathbf{v}_e$ lies in the orthogonal subspace to the vector space generated by $\mathbf{n}$. As a result, given the approximations taken, there is no fluid flow through face $\mathcal{S}_{Pe}$.

The Upwind–Difference Scheme assigns ϕ_e the value of ϕ at the upstream node, that is,

$$\phi_e = \begin{cases} \phi_P & \text{if } (\mathbf{v}\cdot\mathbf{n})_e > 0 \\ \phi_E & \text{if } (\mathbf{v}\cdot\mathbf{n})_e < 0 \end{cases} \tag{3.35}$$

The scheme is summarized in figures 3.3 and 3.4.

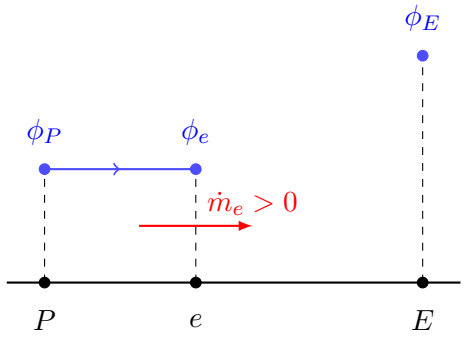


Figure 3.3. UDS when $(\mathbf{v}\cdot\mathbf{n})_e > 0$.

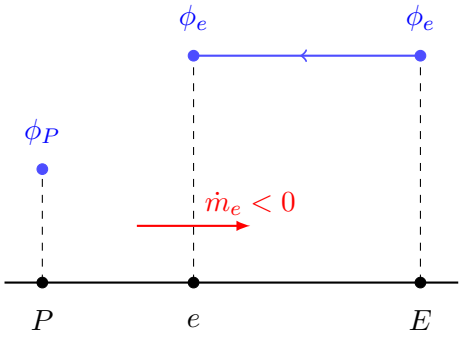


Figure 3.4. UDS when $(\mathbf{v}\cdot\mathbf{n})_e < 0$.

Equation (3.35) can be expressed in a more compact fashion as follows,

$$\dot{m}_e(\phi_e - \phi_P) = \frac{\dot{m}_e - |\dot{m}_e|}{2}(\phi_E - \phi_P) \tag{3.36}$$

since the approximation to compute $\dot{m}_e$ is related to $(\mathbf{v}\cdot\mathbf{n})_e$ through the relation $\dot{m}_e = (\mathbf{v}\cdot\mathbf{n})_e S_{Pe}$. The extension of (3.36) to the remaining faces is the following:

$$\dot{m}_w(\phi_w - \phi_P) = \frac{\dot{m}_w + |\dot{m}_w|}{2}(\phi_W - \phi_P) \tag{3.37}$$

$$\dot{m}_n(\phi_n - \phi_P) = \frac{\dot{m}_n - |\dot{m}_n|}{2}(\phi_N - \phi_P) \tag{3.38}$$

$$\dot{m}_s(\phi_s - \phi_P) = \frac{\dot{m}_s + |\dot{m}_s|}{2}(\phi_S - \phi_P) \tag{3.39}$$

UDS is a stable scheme, however it suffers from numerical diffusion. Indeed, assuming the upstream node is P, expanding ϕ about point x_P in its Taylor expansion up to 2^{nd} degree and using Lagrange's remainder,

$$\phi_e = \phi_P + \left(\frac{\partial\phi}{\partial x}\right)_P d_{Pe} + \left(\frac{\partial^2\phi}{\partial x^2}\right)_{\xi_1} \frac{d_{Pe}^2}{2} \tag{3.40}$$

it is apparent that UDS retains the first term on the left–hand side of (3.40). As a consequence, the error highest order is $(\partial_x\phi)_P d_{Pe}$, which is proportional to the distance between P and the face $\mathcal{S}_{Pe}$. This term resembles to a diffusion flux given, for instance, by Fourier's or Fick's laws of diffusion. The same result is obtained when E is the upstream node,

$$\phi_e = \phi_E - \left(\frac{\partial\phi}{\partial x}\right)_E d_{Ee} + \left(\frac{\partial^2\phi}{\partial x^2}\right)_{\xi_2} \frac{d_{Ee}^2}{2} \tag{3.41}$$

whence it can be deduced that the error is bounded by $\max\{|(\partial_x\phi)_E d_{Pe}|, |(\partial_x\phi)_E d_{Ee}|\}$. The numerical diffusion issue is magnified in multidimensional problems, where peaks of rapid variation can be obtained, hence very fine grids are required.

3.4.2 Central–Difference Scheme (CDS)

The Central–Difference Scheme assumes a linear distribution for ϕ as illustrated in figure 3.5.

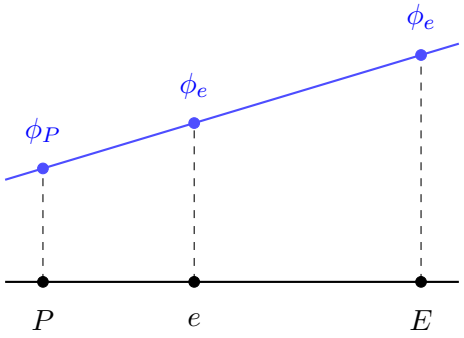


Figure 3.5. Central Difference Scheme (CDS).

Thereby ϕ_e can be obtained interpolating between ϕ_P and ϕ_E,

$$\phi_e - \phi_P = \frac{d_{Pe}}{d_{PE}}(\phi_E - \phi_P) \tag{3.42}$$

This yields a 2$^{\text{nd}}$ order approximation for ϕ_e if $d_{Pe} = d_{Ee}$. In effect, applying Taylor's theorem about point x_e,

$$\phi_P = \phi_e - \left(\frac{\partial\phi}{\partial x}\right)_e d_{Pe} + \frac{1}{2}\left(\frac{\partial^2\phi}{\partial x^2}\right)_e d_{Pe}^2 + \frac{1}{6}\left(\frac{\partial^3\phi}{\partial x^3}\right)_{\xi_1} d_{Pe}^3 \tag{3.43}$$

The 2$^{\text{nd}}$ order approximation of $(\partial_x\phi)_e$ is given by

$$\left(\frac{\partial\phi}{\partial x}\right)_e = \frac{\phi_E - \phi_P}{d_{PE}} - \left(\frac{\partial^3\phi}{\partial x^3}\right)_{\xi_2} \frac{d_{PE}^2}{3!} = \frac{\phi_E - \phi_P}{d_{PE}} - \left(\frac{\partial^3\phi}{\partial x^3}\right)_{\xi_2} \frac{(d_{Pe} + d_{Ee})^2}{3!} \tag{3.44}$$

Introducing (3.44) in (3.43) and imposing $d_{Pe} = d_{Ee}$,

$$\phi_e - \phi_P = \frac{d_{Pe}}{d_{PE}}(\phi_E - \phi_P) - \left(\frac{\partial^2\phi}{\partial x^2}\right)_e \frac{d_{Pe}^2}{2} - \left\{\left(\frac{\partial^3\phi}{\partial x^3}\right)_{\xi_1} + 4\left(\frac{\partial^3\phi}{\partial x^3}\right)_{\xi_2}\right\} \frac{d_{Pe}^3}{6} \tag{3.45}$$

As CDS retains the first term on the left–hand side of (3.45), the highest order term of the error is $\frac{1}{2}(\partial_x^2\phi)_e d_{Pe}^2$, proving that CDS provides a 2$^{\text{nd}}$ order approximation of ϕ_e when $d_{Pe} = d_{Ee}$. Nonetheless, this scheme is prone to stability problems producing oscillatory outputs since the approximation is of order higher than 1.

3.4.3 Exponential–Difference Scheme (EDS)

The exponential difference scheme assumes a distribution for ϕ based on the steady 2–dimensional generalized convection–diffusion equation with no source term, that is to say,

$$\frac{\mathrm{d}}{\mathrm{d}x}(\rho u\phi) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\Gamma\frac{\mathrm{d}\phi}{\mathrm{d}x}\right) \tag{3.46}$$

So as to ease the study, ρu and Γ are assumed to be constant. Thereby the initial value problem obtained is

$$\begin{cases} \dfrac{\mathrm{d}^2\phi}{\phi x^2} - \dfrac{\rho u}{\Gamma}\dfrac{\mathrm{d}\phi}{\mathrm{d}x} = 0 \quad \text{in } (x_P, x_E) \subset \mathbb{R} \\ \phi(x_P) = \phi_P \\ \phi(x_E) = \phi_E \end{cases} \tag{3.47}$$

Since the initial value problem (3.47) involves a second order linear ODE with two boundary conditions, its solutions exists and is unique due to theorem C.13, and is given by

$$\phi(x) = \phi_P + \frac{e^{\frac{\rho u}{\Gamma}(x - x_P)} - 1}{e^{\frac{\rho u}{\Gamma} d_{PE}} - 1}(\phi_E - \phi_P) \tag{3.48}$$

Péclet's number is defined as the ratio between of strengths of convection and diffusion [6],

$$\mathrm{Pe} = \frac{\text{convection transport rate}}{\text{diffusion transport rate}} = \frac{\rho u L}{\Gamma} \tag{3.49}$$

where L is a characteristic length of the problem. Taking d_{PE} as characteristic length and evaluating (3.48) at $x = x_e$, the approximation of ϕ_e given by EDS in terms of Péclet's number is written as

$$\phi_e - \phi_P = \frac{e^{\mathrm{Pe}_e \frac{d_{Pe}}{d_{PE}}} - 1}{e^{\mathrm{Pe}_e} - 1}(\phi_E - \phi_P) \tag{3.50}$$

3.4.4 Second–order Upwind Linear Extrapolation (SUDS)

As previously mentioned, incompressible flows and fluids at low Mach number are more influenced by upstream condition than by downstream conditions. In order to account for this fact and to ease the study, we introduce a new notation. When located at the face separating two control volumes, f refers to the face, D is the downstream node, C is the first upstream node and U is the most upstream node. Some books may use U and UU instead of C and U, respectively.

The Second–order Upwind Linear Extrapolation scheme takes profit of this idea since it extrapolates ϕ_e using a straight line between the values of ϕ at nodes C and U. The two possible situations are pictured in figures 3.6 and 3.7.

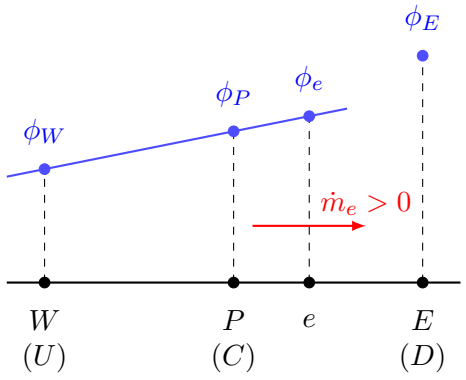


Figure 3.6. SUDS when $(\mathbf{v} \cdot \mathbf{n})_e > 0$.

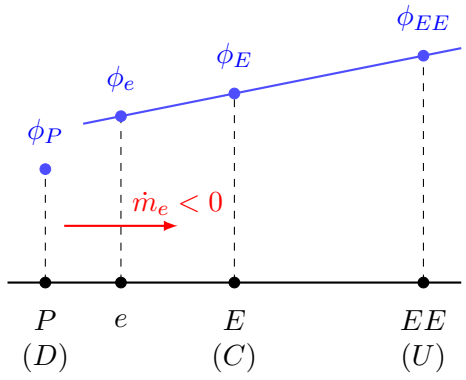


Figure 3.7. SUDS when $(\mathbf{v} \cdot \mathbf{n})_e < 0$.

On the one hand, when $(\mathbf{v} \cdot \mathbf{n})_e > 0$, the line between points (x_W, ϕ_W) and (x_P, ϕ_P) is given by

$$\phi(x) = \phi_W + \frac{\phi_P - \phi_W}{d_{PW}}(x - x_W) \tag{3.51}$$

and substituting at $x = x_e$, the formula for ϕ_e is obtained:

$$\phi_e = \phi_W + \frac{\phi_P - \phi_W}{d_{PW}}(x_e - x_W) = \phi_P + \frac{d_{Pe}}{d_{PW}}(\phi_P - \phi_W) \tag{3.52}$$

On the other hand, in the case of $(\mathbf{v} \cdot \mathbf{n})_e < 0$, the line between points (x_E, ϕ_E) and (x_{EE}, ϕ_{EE}) is

$$\phi(x) = \phi_E + \frac{\phi_{EE} - \phi_E}{d_{E,EE}}(x - x_E) \tag{3.53}$$

and the approximation of ϕ_e is

$$\phi_e = \phi_E + \frac{\phi_{EE} - \phi_E}{d_{E,EE}}(x_e - x_E) = \phi_E + \frac{d_{Ee}}{d_{E,EE}}(\phi_E - \phi_{EE}) \tag{3.54}$$

Using the DCU notation, (3.52) and (3.54) are both rewritten in the following manner:

$$\phi_f - \phi_C = \frac{d_{Cf}}{d_{CU}}(\phi_C - \phi_U) \tag{3.55}$$

In order to prove that SUDS is a second order scheme when a uniform mesh is used and $(\mathbf{v} \cdot \mathbf{n})_e > 0$, consider the Taylor expansion up to 2^{nd} degree of ϕ about point x_W,

$$\phi_e = \phi_W + \left(\frac{\partial \phi}{\partial x}\right)_W d_{We} + \left(\frac{\partial^2 \phi}{\partial x^2}\right)_{\xi_1} \frac{d_{We}^2}{2} \tag{3.56}$$

The first derivative of ϕ with respect to x can be replaced by its first order approximation, namely,

$$\left(\frac{\partial \phi}{\partial x}\right)_W = \frac{\phi_P - \phi_W}{d_{PW}} - \left(\frac{\partial^2 \phi}{\partial x^2}\right)_{\xi_2} \frac{d_{PW}}{2} \tag{3.57}$$

thereby,

$$\begin{aligned} \phi_e &= \phi_W + \frac{d_{We}}{d_{PW}}(\phi_P - \phi_W) + \left(\frac{\partial^2 \phi}{\partial x^2}\right)_{\xi_1} \frac{d_{We}^2}{2} - \left(\frac{\partial^2 \phi}{\partial x^2}\right)_{\xi_2} \frac{d_{We} d_{PW}}{2} \\ &= \phi_P + \frac{d_{Pe}}{d_{PW}}(\phi_P - \phi_W) + \left(\frac{\partial^2 \phi}{\partial x^2}\right)_{\xi_1} \frac{(d_{PW} + d_{Pe})^2}{2} - \left(\frac{\partial^2 \phi}{\partial x^2}\right)_{\xi_2} \frac{(d_{PW} + d_{Pe}) d_{PW}}{2} \end{aligned} \tag{3.58}$$

The scheme retains the two first terms on the right–hand side of (3.58), therefore the error is composed by the last two terms. The uniform mesh hypothesis implies $d_{PW} = 2d_{Pe} = L$, therefore the error term is multiplied by L^2,

$$\phi_e = \phi_P + \frac{d_{Pe}}{d_{PW}}(\phi_P - \phi_W) + \frac{3L^2}{4}\left\{3\left(\frac{\partial^2 \phi}{\partial x^2}\right)_{\xi_1} - \left(\frac{\partial^2 \phi}{\partial x^2}\right)_{\xi_2}\right\} \tag{3.59}$$

whence the second order of SUDS is deduced. The proof in the case of $(\mathbf{v} \cdot \mathbf{n})_e < 0$ is analogous.

3.4.5 Quadratic Upwind Interpolation for Convective Kinematics (QUICK)

A logical improvement of CDS is using a parabola to interpolate between nodal points rather than a straight line. To construct a parabola three points are needed. As aforementioned, upstream conditions

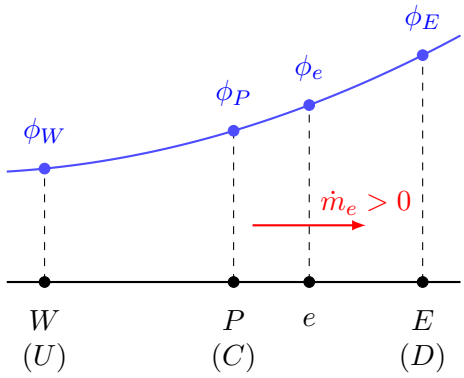


Figure 3.8. QUICK when $(\mathbf{v}\cdot\mathbf{n})_e > 0$.

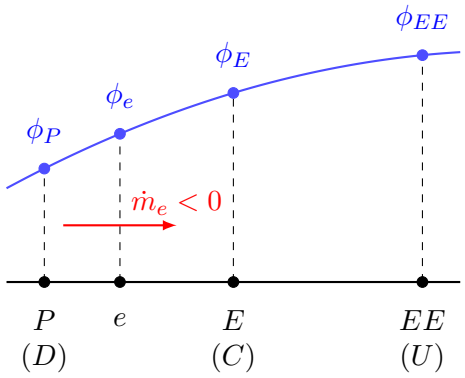


Figure 3.9. QUICK when $(\mathbf{v}\cdot\mathbf{n})_e < 0$.

have a greater influence on flow properties than downstream conditions for incompressible flows and low Mach number gases. QUICK scheme takes profit of this fact.

Let (x_0,ϕ_0), (x_1,ϕ_1), (x_2,ϕ_2) be the points which the polynomial $p(x)$ must interpolate, that is, $p(x_0)=\phi_0$, $p(x_1)=\phi_1$ and $p(x_2)=\phi_2$, satisfying $x_0<x_1<x_2$. If $(\mathbf{v}\cdot\mathbf{n})_e>0$ then $x_0=x_W$, $x_1=x_P$ and $x_2=x_E$, whereas $x_0=x_P$, $x_1=x_E$ and $x_2=x_{EE}$ in the case of $(\mathbf{v}\cdot\mathbf{n})_e<0$. Let $p(x)$ be the following polynomial

$$p(x)=a_0+a_1(x-x_0)+a_2(x-x_0)(x-x_1),\quad a_0,a_1,a_2\in\mathbb{R} \tag{3.60}$$

Since the interpolating polynomial exists and is unique (see [9], Theorem 8.1), by imposing the interpolating condition, $p(x)$ will be the desired polynomial. The interpolating condition is,

$$\left.\begin{aligned} p(x_0)&=a_0=\phi_0\\ p(x_1)&=a_0+a_1(x_1-x_0)=\phi_1\\ p(x_2)&=a_0+a_1(x_2-x_0)+a_2(x_2-x_0)(x_2-x_1)=\phi_2 \end{aligned}\right\} \tag{3.61}$$

which yields the following linear system:

$$\begin{pmatrix} 1 & 0 & 0\\ 1 & x_1-x_0 & 0\\ 1 & x_2-x_0 & (x_2-x_1)(x_2-x_0)\end{pmatrix}\begin{pmatrix}a_0\\a_1\\a_2\end{pmatrix}=\begin{pmatrix}\phi_0\\\phi_1\\\phi_2\end{pmatrix} \tag{3.62}$$

The determinant of the system matrix is non-zero because the abscissae are distinct, therefore the solution is given by

$$\left.\begin{aligned} a_0&=\phi_0\\ a_1&=\frac{\phi_1-\phi_0}{x_1-x_0}\\ a_2&=\frac{\phi_2-\phi_0}{(x_2-x_1)(x_2-x_0)}-\frac{\phi_1-\phi_0}{(x_2-x_1)(x_1-x_0)} \end{aligned}\right\} \tag{3.63}$$

and the polynomial is

$$p(x)=\phi_0-\frac{(x-x_2)(x-x_0)}{(x_2-x_1)(x_1-x_0)}(\phi_1-\phi_0)+\frac{(x-x_1)(x-x_0)}{(x_2-x_1)(x_2-x_0)}(\phi_2-\phi_0) \tag{3.64}$$

In terms of the *DCU* notation, we have the following:

$$p(x) = \begin{cases} \phi_U - \dfrac{(x-x_D)(x-x_U)}{(x_D-x_C)(x_C-x_U)}(\phi_C-\phi_U) + \dfrac{(x-x_C)(x-x_U)}{(x_D-x_C)(x_D-x_U)}(\phi_D-\phi_U) & \text{if } (\mathbf{v}\cdot\mathbf{n})_e > 0 \\ \phi_D - \dfrac{(x-x_U)(x-x_D)}{(x_U-x_C)(x_C-x_D)}(\phi_C-\phi_D) + \dfrac{(x-x_C)(x-x_D)}{(x_U-x_C)(x_U-x_D)}(\phi_U-\phi_D) & \text{if } (\mathbf{v}\cdot\mathbf{n})_e < 0 \end{cases} \tag{3.65}$$

Assuming a uniform grid, i.e. $x_1 - x_0 = x_2 - x_1 = L$ and the face f located at the midpoint between nodal points, the approximation of ϕ_e given by QUICK scheme is

$$\phi_e = -\frac{1}{8}\phi_0 + \frac{6}{8}\phi_1 + \frac{3}{8}\phi_2 \tag{3.66}$$

and depending on the sign of $(\mathbf{v}\cdot\mathbf{n})_e$,

$$\phi_e = \begin{cases} -\dfrac{1}{8}\phi_U + \dfrac{6}{8}\phi_C + \dfrac{3}{8}\phi_D & \text{if } (\mathbf{v}\cdot\mathbf{n})_e > 0 \\ -\dfrac{1}{8}\phi_D + \dfrac{6}{8}\phi_C + \dfrac{3}{8}\phi_U & \text{if } (\mathbf{v}\cdot\mathbf{n})_e < 0 \end{cases} \tag{3.67}$$

The output (3.67) provided by QUICK scheme is second–order accurate.

3.4.6 Normalization of variables

Owing to numerical reasons, it is convenient to normalize spatial and convective variables, that is to say, define new variables which take a rather small range of values. This is accomplished using the *DCU* notation and defining

$$\hat{x} = \frac{x - x_U}{x_D - x_U} \tag{3.68}$$

$$\hat{\phi} = \frac{\phi - \phi_U}{\phi_D - \phi_U} \tag{3.69}$$

Of course, $(\hat{x}_U, \hat{\phi}_U) = (0,0)$, $(\hat{x}_D, \hat{\phi}_D) = (1,1)$ and $\hat{x}_C, \hat{x}_f \in [0,1]$. However, $\hat{\phi}$ is not necessarily in $[0,1]$ for all $x \in [0,1]$, nor does it have to be an increasing function. These situations are represented in figures 3.10 and 3.11.

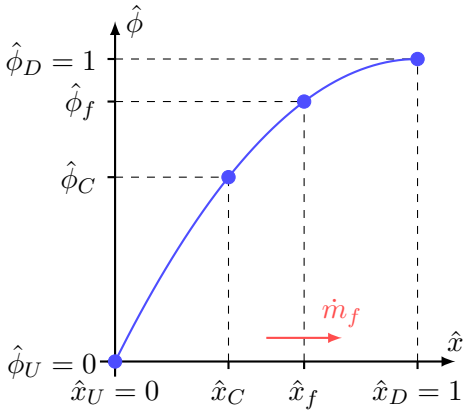


Figure 3.10. Scheme of normalized variables when $\hat{\phi}(x)$ is a strictly increasing function.

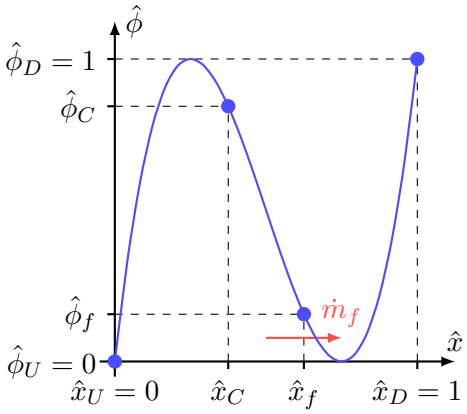


Figure 3.11. Scheme of normalized variables when $\hat{\phi}(x)$ is not a strictly increasing function.

Some schemes, such as SMART, give the value of the normalized variable at face $\hat{\phi}_f$ directly as equation (3.71) shows. Based on $\hat{\phi}_f$, the variable at face is calculated by

$$\phi_f = \phi_U + \hat{\phi}_f(\phi_D - \phi_U) \tag{3.70}$$

3.4.7 Sharp and Monotonic Algorithm for Realistic Transport (SMART)

As aforementioned, schemes whose order is higher than one might be unstable, producing oscillatory outputs for the convective variables. For instance, CDS, SUDS and QUICK are not bounded schemes. The conditions for stability and accuracy are formulated in [10]:

(i) $\hat{\phi}_f$ must be a continuous function of $\hat{\phi}_C$.

(ii) If $\hat{\phi}_C = 0$, then $\hat{\phi}_f = 0$.

(iii) If $\hat{\phi}_C = 1$, then $\hat{\phi}_f = 1$.

(iv) If $0 < \hat{\phi}_f < 1$, then $\hat{\phi}_C < \hat{\phi}_f < 1$.

Conditions (i) through (iv) are represented in figure 3.12. A bounded convective scheme must output results lying within the shadowed region.

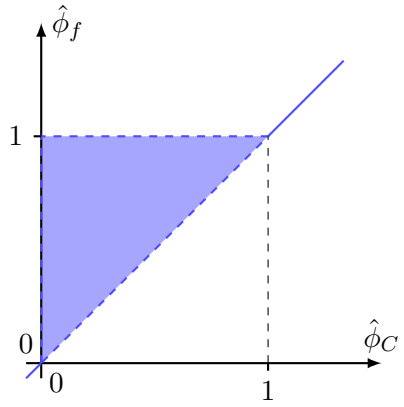


Figure 3.12. High–order bounded convection schemes conditions for stability.

The SMART scheme (Sharp and Monotonic Algorithm for Realistic Transport) is a bounded convective scheme [10], given by:

$$\hat{\phi}_f = \begin{cases} -\dfrac{\hat{x}_f(1-3\hat{x}_C+2\hat{x}_f)}{\hat{x}_C(\hat{x}_C-1)}\hat{\phi}_C & \text{if} \quad 0 < \hat{\phi}_C < \dfrac{\hat{x}_C}{3} \\ \dfrac{\hat{x}_f(\hat{x}_f-\hat{x}_C)}{1-\hat{x}_C} + \dfrac{\hat{x}_f(\hat{x}_f-1)}{\hat{x}_C(\hat{x}_C-1)}\hat{\phi}_C & \text{if} \quad \dfrac{\hat{x}_C}{3} < \hat{\phi}_C < \dfrac{\hat{x}_C(1+\hat{x}_f-\hat{x}_C)}{\hat{x}_f} \\ 1 & \text{if} \quad \dfrac{\hat{x}_C(1+\hat{x}_f-\hat{x}_C)}{\hat{x}_f} < \hat{\phi}_C < 1 \\ \hat{\phi}_C & \text{otherwise} \end{cases} \tag{3.71}$$

3.5 Final form of the generalized convection–diffusion equation

The purpose of this subsection is to obtain a discretization equation of the form

$$\mathcal{A}_P\phi_P + \sum_F \mathcal{A}_F\phi_F = \mathcal{Q}_P \tag{3.72}$$

so that it can be easily implemented to be solved numerically, starting from equation (3.32) and the studied schemes to evaluate convective properties. Among the revised schemes, some use only the surrounding nodes, whilst others involve a larger amount of nodes. As a consequence of the different treatment needed, separate subsections are devoted to each type of scheme.

3.5.1 Small molecule schemes

Small molecule schemes are those which only involve adjacent nodes to the volume faces, i.e. the subindex F in (3.72) refers to nodes E, W, N and S. For instance, UDS, CDS and EDS are small molecule schemes. As a result, small molecule schemes can be introduced in a compact form, nonetheless we shall not repeat the entire discussion here. The whole development can be found at [6].

Recall that the mass flow rates through the faces of $\mathcal{V}_P$ are calculated as

$$\dot{m}_e = (\rho u)_e\, S_e, \quad \dot{m}_w = (\rho u)_w\, S_w, \quad \dot{m}_n = (\rho v)_s\, S_n, \quad \dot{m}_s = (\rho v)_s\, S_s \tag{3.73}$$

In equation (3.24) we defined D_f, which particularized for each face results in the following coefficients:

$$D_e = \frac{\Gamma_e S_e}{d_{PE}}, \quad D_w = \frac{\Gamma_w S_w}{d_{PW}}, \quad D_n = \frac{\Gamma_n S_n}{d_{PN}}, \quad D_s = \frac{\Gamma_s S_s}{d_{PS}} \tag{3.74}$$

Using $\dot{m}_f$ and D_f, Péclet's numbers at faces can be computed as follows:

$$\mathrm{Pe}_e = \frac{F_e}{D_e}, \quad \mathrm{Pe}_w = \frac{F_w}{D_w}, \quad \mathrm{Pe}_n = \frac{F_n}{D_n}, \quad \mathrm{Pe}_s = \frac{F_s}{D_s} \tag{3.75}$$

In addition, we define the operator $[\![\cdot,\cdot]\!] : \mathbb{R}^2 \to \mathbb{R}$ as $[\![x, y]\!] = \max\{x, y\}$. According to Patankar, the discretized version of the generalized convection–diffusion equation (3.33) can be transformed into

$$a_P\phi_P = a_E\phi_E + a_W\phi_W + a_N\phi_N + a_S\phi_S + b_P \tag{3.76}$$

where the coefficients are given by

$$a_E = D_e\, A(|\mathrm{Pe}_e|) + [\![-F_e, 0]\!] \tag{3.77}$$
$$a_W = D_w A(|\mathrm{Pe}_w|) + [\![F_w, 0]\!] \tag{3.78}$$
$$a_N = D_n A(|\mathrm{Pe}_n|) + [\![-F_n, 0]\!] \tag{3.79}$$
$$a_S = D_s A(|\mathrm{Pe}_s|) + [\![F_s, 0]\!] \tag{3.80}$$
$$b_P = S_C^\phi V_P + \frac{\rho_P^0 \phi_P^0}{\Delta t} V_P \tag{3.81}$$
$$a_P = a_E + a_W + a_N + a_S + \frac{\rho_P^0 V_P}{\Delta t} - S_P^\phi V_P \tag{3.82}$$

and $A\colon \mathbb{R} \to \mathbb{R}$ is a function which depends upon the chosen scheme. Table 3.1 shows $A(|\mathrm{Pe}|)$ for several schemes. It also includes the Hybrid and Power law schemes which we have not studied.

Scheme	$A(\|\mathrm{Pe}\|)$
Upwind–Difference Scheme	1
Central–Difference Scheme	$1-0.5\|\mathrm{Pe}\|$
Exponential–Difference Scheme	$\|\mathrm{Pe}\|/\left(\exp\left(\|\mathrm{Pe}\|\right)-1\right)$
Hybrid Scheme	$[\![0, 1-0.5\|\mathrm{Pe}\|]\!]$
Power law Scheme	$\left[\!\left[0, \left(1-0.5\|\mathrm{Pe}\|\right)^5\right]\!\right]$

Table 3.1. Function $A(|\mathrm{Pe}|)$ for different schemes[6].

3.5.2 Large molecule schemes

High–resolution schemes (HRS) such as SUDS, QUICK and SMART, not only use adjacent nodes to the faces but also the most upstream nodes, that is to say, involve a larger molecule. Since a larger molecule increases the memory usage and the computational effort, it is desirable to keep it as low as possible. Therefore, the aim is to obtain a discretization equation such as (3.76), where only the surrounding nodes participate, while upstream nodes are computed by different means and collected in b_P.

The first logical solution would be to use small molecule schemes, although it must be kept in mind the lower order of the approximations. The second solution would be to compute the upstream node value using the data of the previous iteration and introduce this term in the equation as a contribution to b_P. Nevertheless, this may lead to the divergence of the iterations since the terms treated explicitly may be substantial [11].

The solution is to compute the approximated terms with a higher order approximation explicitly and put them on the right–hand side of equation (3.72). Then a simpler approximation to these terms, for instance one that provides a smaller molecule, is put on the left–hand side and on the right–hand side, computing it using explicit values. Then the right–hand side is the difference between two approximations of the same value, hence is likely to be small. This technique is known as deferred correction, and is used with higher–order approximations, as well as grid non–orthogonality and correction to prevent undesirable effects in solutions [11].

Given a face f, the idea is approximate ϕ_f as

$$\phi_f^{\mathrm{HRS}} - \phi_P = (\phi_f^{\mathrm{UDS}} - \phi_P) + (\phi_f^{\mathrm{HRS},*} - \phi_f^{\mathrm{UDS},*}) \tag{3.83}$$

ϕ_f^{HRS} and ϕ_f^{UDS} are the current calculated values of ϕ using the chosen HRS and UDS, whereas $\phi_f^{\mathrm{HRS},*}$ and $\phi_f^{\mathrm{UDS},*}$ are the computed values in the previous iteration. As stated above, when convergence is achieved, $\phi_f^{\mathrm{HRS}} = \phi_f^{\mathrm{HRS},*}$ and $\phi_f^{\mathrm{UDS}} = \phi_f^{\mathrm{UDS},*}$ [4]. Substituting $\phi_f - \phi_P$ by $\phi_f^{\mathrm{HRS}} - \phi_P$ in (3.31)

$$\begin{aligned}\rho_P^0 \frac{\phi_P - \phi_P^0}{\Delta t} V_P + \dot{m}_e(\phi_e^{\mathrm{HRS}} - \phi_P) - \dot{m}_w(\phi_w^{\mathrm{HRS}} - \phi_P) + \dot{m}_n(\phi_n^{\mathrm{HRS}} - \phi_P) - \dot{m}_s(\phi_s^{\mathrm{HRS}} - \phi_P) = \\ = D_e(\phi_E - \phi_P) - D_w(\phi_P - \phi_W) + D_n(\phi_N - \phi_P) - D_s(\phi_P - \phi_S) + (S_C^\phi + S_P^\phi \phi_P)V_P\end{aligned} \tag{3.84}$$

and using relation (3.83)

$$\begin{aligned}\rho_P^0\frac{\phi_P-\phi_P^0}{\Delta t}V_P+\dot{m}_e(\phi_e^{\text{UDS}}-\phi_P)-\dot{m}_w(\phi_w^{\text{UDS}}-\phi_P)+\dot{m}_n(\phi_n^{\text{UDS}}-\phi_P)-\dot{m}_s(\phi_s^{\text{UDS}}-\phi_P)=\\ =D_e(\phi_E-\phi_P)-D_w(\phi_P-\phi_W)+D_n(\phi_N-\phi_P)-D_s(\phi_P-\phi_S)+(S_C^\phi+S_P^\phi\phi_P)V_P+\\ -\dot{m}_e(\phi_e^{\text{HRS},*}-\phi_e^{\text{UDS},*})+\dot{m}_w(\phi_w^{\text{HRS},*}-\phi_w^{\text{UDS},*})-\dot{m}_n(\phi_n^{\text{HRS},*}-\phi_n^{\text{UDS},*})+\dot{m}_s(\phi_s^{\text{HRS},*}-\phi_s^{\text{UDS},*})\end{aligned} \tag{3.85}$$

Replacing the corresponding terms with expressions (3.36) through (3.39) and rearranging terms, the desired expression is found

$$a_P\phi_P=a_W\phi_W+a_E\phi_E+a_S\phi_S+a_N\phi_N+b_P \tag{3.86}$$

with the following coefficients:

$$a_E=D_e-\frac{\dot{m}_e-|\dot{m}_e|}{2}=\frac{\Gamma_eS_e}{d_{PE}}-\frac{\dot{m}_e-|\dot{m}_e|}{2} \tag{3.87}$$

$$a_W=D_w+\frac{\dot{m}_w+|\dot{m}_w|}{2}=\frac{\Gamma_wS_w}{d_{PW}}+\frac{\dot{m}_w+|\dot{m}_w|}{2} \tag{3.88}$$

$$a_N=D_n-\frac{\dot{m}_n-|\dot{m}_n|}{2}=\frac{\Gamma_nS_n}{d_{PN}}-\frac{\dot{m}_n-|\dot{m}_n|}{2} \tag{3.89}$$

$$a_S=D_s+\frac{\dot{m}_s+|\dot{m}_s|}{2}=\frac{\Gamma_sS_s}{d_{PS}}+\frac{\dot{m}_s+|\dot{m}_s|}{2} \tag{3.90}$$

$$a_P=a_W+a_E+a_S+a_N+\frac{\rho_P^0V_P}{\Delta t}-S_P^\phi V_P \tag{3.91}$$

$$\begin{aligned}b_P=\frac{\rho_P^0\phi_P^0}{\Delta t}V_P+S_C^\phi V_P-\dot{m}_e(\phi_e^{\text{HRS},*}-\phi_e^{\text{UDS},*})+\dot{m}_w(\phi_w^{\text{HRS},*}-\phi_w^{\text{UDS},*})\\ -\dot{m}_n(\phi_n^{\text{HRS},*}-\phi_n^{\text{UDS},*})+\dot{m}_s(\phi_s^{\text{HRS},*}-\phi_s^{\text{UDS},*})\end{aligned} \tag{3.92}$$

3.6 Treatment of boundary conditions

In Cauchy problems involving Partial Differential Equations (PDEs), there exist several kinds of boundary conditions which must be prescribed in order to guarantee the existence and uniqueness of solution, although in this project we will only consider two types. So as to illustrate how these conditions are set, let $U\subset\mathbb{R}^n$ be a bounded open subset of $\mathbb{R}^n$. The heat or diffusion equation is the PDE

$$u_t-\Delta u=f(x,t)\quad(x,t)\in U\times(0,\infty) \tag{3.93}$$

where $\Delta=\sum_{i=1}^m\frac{\partial^2}{\partial x_i^2}$ is Laplace's operator and f models the internal sources for magnitude u [5]. This equation models the evolution in time of the density u of some quantity such as heat, chemical concentration, etc. Let $g\colon U\to\mathbb{R}$ be the initial value for u. The typical Cauchy problem for diffusion equation is

$$\begin{cases}u_t-\Delta u=f(x,t) & \text{in } U\times(0,\infty)\\ u=g & \text{on } U\times\{t=0\}\\ \text{Boundary conditions}\end{cases} \tag{3.94}$$

The boundary conditions considered are:

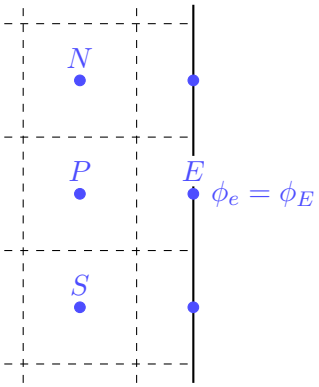


Figure 3.13. Dirichlet boundary condition.

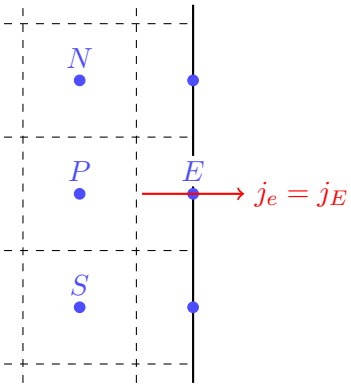


Figure 3.14. Neumann boundary condition.

- Dirichlet boundary condition: the value of u is prescribed on $\partial U \times (0,\infty)$, that is to say, if $d\colon \partial U \times (0,\infty) \to \mathbb{R}$, $(x,t) \mapsto d(x,t)$ describes the boundary condition, then (3.94) is written as

$$\begin{cases} u_t - \Delta u = f(x,t) & \text{in } U \times (0,\infty) \\ u = g & \text{on } U \times \{t=0\} \\ u = d & \text{on } \partial U \times (0,\infty) \end{cases} \tag{3.95}$$

  When (3.93) is thought of as describing the propagation of heat, then d fixes the temperature at the boundary of U for each time.

- Neumann boundary condition: the normal derivative of u to the boundary of U is prescribed on $\partial U \times (0,\infty)$, i.e. if $n\colon \partial U \times (0,\infty) \to \mathbb{R}$ describes the boundary condition, then (3.94) is written as

$$\begin{cases} u_t - \Delta u = f(x,t) & \text{in } U \times (0,\infty) \\ u = g & \text{on } U \times \{t=0\} \\ \partial_\nu u = n & \text{on } \partial U \times (0,\infty) \end{cases} \tag{3.96}$$

  where ν is the outer normal vector to ∂U. In terms of heat, this boundary condition sets the conduction heat transfer through U for each time.

The numerical treatment of boundary conditions is straightforward, specially is our case as we are using a cartesian mesh on a rectangular domain. In the case of a Dirichlet boundary condition, such as the one shown in figure 3.13, the value at face must be equal to the prescribed value at boundary, that is,

$$\phi_e = \phi_E \tag{3.97}$$

and flux per unit of surface can be easily computed as

$$j_e = -\Gamma_P \frac{\phi_e - \phi_P}{d_{Pe}} \tag{3.98}$$

In contrast, when a Neumann boundary condition with flux j_e is imposed, the value at face is

$$\phi_e = \phi_P - \frac{j_e d_{Pe}}{\Gamma_P} \tag{3.99}$$

This second situation is pictured in figure 3.14.

3.7 Solving algorithm

The procedure to solve of a transient convection–diffusion problem with 2D–cartesian mesh is shown in Algorithm 1.

Algorithm 1 Resolution of a transient convection–diffusion problem with 2D–cartesian mesh.

1 Input data:

1.1 Physical data: geometry, thermophysical properties, initial and boundary conditions.

1.2 Numerical data: mesh, Δt (time step), δ (convergence criterion).

2 Mesh generation: nodes position, faces position, distances, surfaces and volumes.

3 Initial map: $n \leftarrow 0$, $t^n \leftarrow 0$, $\phi^0[i][j] = \phi(x, y, t)|_{t=0}$.

4 Compute the new time step: $t^{n+1} = t^n + \Delta t$.

4.1 Initial estimated values: $\phi^*[i][j] \leftarrow \phi^0[i][j]$.

4.2 Evaluation of the discretization coefficients: $a_E[i][j]$, $a_W[i][j]$, $a_N[i][j]$, $a_S[i][j]$, $a_P[i][j]$, $b_P[i][j]$.

4.3 Resolution of the linear system

$$a_P[i][j]\,\phi[i][j] = a_E[i][j]\,\phi[i+1][j] + a_W[i][j]\,\phi[i-1][j] + a_N[i][j]\,\phi[i][j+1] + a_S[i][j]\,\phi[i][j-1] + b_P[i][j]$$

4.4 Is $\max_{i,j} |\phi^*[i][j] - \phi[i][j]| < \delta$?

- Yes: continue.
- No: $\phi^*[i][j] \leftarrow \phi[i][j]$, go to **4.2**.

5 New time step?

- Yes: $n \leftarrow n + 1$, go to **4**.
- No: continue.

6 Final computations, print results.

7 End.

Hereinafter, the term iteration will be used to refer to the iterative procedure to solve the linear system on step **4.3**. It must not be confused with the next time instant term.

On step **4** the next time instant is computed. This is the most computationally expensive step in the algorithm, specially part **4.3** where the resolution of the linear system of discretized equations is carried out. As a result of the convection–diffusion equations nature, the system matrix A and the vector of independent terms b change each time the convergence condition is not fulfilled on step (**4.4**). Since A and b depend on the previous iteration value of ϕ, that is to say, $A = A(\phi^*)$ and $b = b(\phi^*)$, the linear system of equations is

$$A(\phi^*)\,\phi = b(\phi^*) \tag{3.100}$$

In the case both A and b were constant, the algorithm needed to solve the linear system (3.100) would be clear at first glance. By looking at equations (3.76) and (3.86) (actually the same equation), the value of ϕ_P is only influenced by ϕ_E, ϕ_W, ϕ_S and ϕ_N, hence A is a pentadiagonal by blocks matrix, therefore A is sparse, i.e. most of the elements are zero. In this situation an iterative method for solving the linear system is convenient.

Let A_{ij} denote the element in the i-th row and j-th column of A. Assume A is a strictly diagonally

dominant (SDD) matrix, that is to say,

$$|A_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |A_{ij}|, \quad 1 \leq i \leq n \tag{3.101}$$

where n is the dimension of the matrix. Then the Gauss–Seidel algorithm is guaranteed to converge and eventually solve the system. In terms of the discretization coefficients, condition (3.101) is written as

$$|a_P[i][j]| \geq |a_W[i][j]| + |a_E[i][j]| + |a_S[i][j]| + |a_N[i][j]| \tag{3.102}$$

Recall that in section 3.3 we linearized the linear term as $\dot{\bar{s}}_\phi = S_C^\phi + S_P^\phi \phi_P$ and we asked S_P^ϕ to be non–positive, i.e., $S_P^\phi \leq 0$. By looking at coefficient a_P for small molecule schemes given by (3.82), we notice the following:

$$a_P = a_E + a_W + a_N + a_S + \frac{\rho_P^0 V_P}{\Delta t} - S_P^\phi V_P \geq a_E + a_W + a_N + a_S + \frac{\rho_P^0 V_P}{\Delta t} \tag{3.103}$$

Therefore the fact that $S_P^\phi \leq 0$, although not being a sufficient conditions, helps the matrix A to satisfy the SDD condition. In the case an iterative procedure diverges, a direct method to solve the linear system might be the most convenient option. Two methods for solving linear systems are discussed in appendix D.

4 Diagonal flow case

4.1 Statement

Let $L > 0$ be a constant and consider the square domain $\Omega = (0, L) \times (0, L) \subset \mathbb{R}^2$. In Ω we consider the steady state version of the generalized convection–diffusion equation (1.27), with no source term, constant density and constant diffusion coefficient, that is to say,

$$\frac{\rho}{\Gamma}\mathbf{v} \cdot \boldsymbol{\nabla}\phi = \Delta\phi \tag{4.1}$$

The following Dirichlet boundary conditions are prescribed:

- $\phi = \phi_{\text{low}}$ on $C_1 = [0, L) \times \{0\} \cup \{L\} \times [0, L)$.
- $\phi = \phi_{\text{high}}$ on $C_2 = \{0\} \times (0, L] \cup (0, L] \times \{L\}$.

Notice that $C_1,\ C_2 \subset \mathbb{R}^2$ constitute a partition of the boundary of Ω. In order to encode the boundary conditions more easily, we define the function $g\colon \Omega \to \mathbb{R}$ in the following way:

$$g(x, y) = \begin{cases} \phi_{\text{low}} & \text{if } (x, y) \in C_1 \\ \phi_{\text{high}} & \text{if } (x, y) \in C_2 \end{cases} \tag{4.2}$$

The velocity field is $\mathbf{v} = v_0 \cos(\alpha)\mathbf{i} + v_0 \sin(\alpha)\mathbf{j}$ with $v_0 > 0$ constant and $\alpha = \pi/4$, whence

$$\frac{\rho}{\Gamma}\mathbf{v} \cdot \boldsymbol{\nabla}\phi = \frac{\rho v_0 \cos(\alpha)}{\Gamma}(\phi_x + \phi_y) = \underbrace{\frac{\cos(\alpha)}{L}}_{\beta}\underbrace{\frac{\rho v_0 L}{\Gamma}}_{\text{Pe}}(\phi_x + \phi_y) = (\phi_x + \phi_y)\,\beta\,\text{Pe} \tag{4.3}$$

The resulting Cauchy problem is gathered in (4.4) and summarized in figure 4.1.

$$\begin{cases} \Delta\phi - (\phi_x + \phi_y)\,\beta\,\text{Pe} = 0 & \text{in } \Omega \\ \phi = g & \text{on } \partial\Omega \end{cases} \tag{4.4}$$

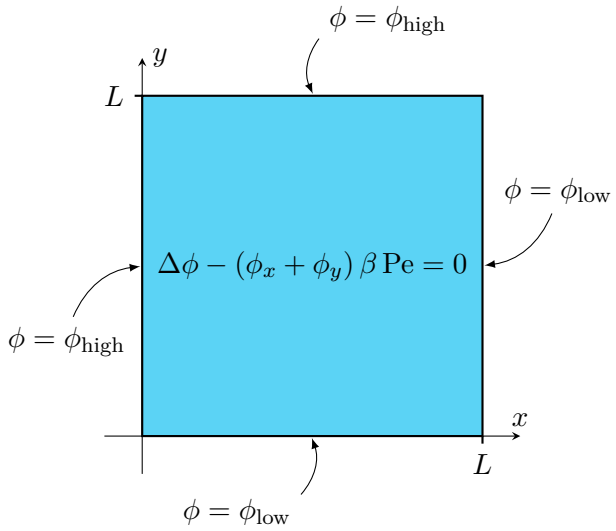


Figure 4.1. Cauchy problem for the diagonal flow case.

4.2 Analytical solution

As we have previously seen, Péclet's number is defined as

$$\mathrm{Pe} = \frac{\text{convection transport rate}}{\text{diffusion transport rate}} = \frac{\rho v_0 L}{\Gamma} \tag{4.5}$$

Note that the factor β in the PDE from problem (4.4) is a constant determined by the geometry, whereas Peclet's number depends on the fluid and on the velocity field. Since no more factors intervene on the PDE, this tells us that the behaviour of the solution will depend greatly on Peclet's number.

4.2.1 Classical solution for $\mathrm{Pe} = \infty$

Whenever $\mathrm{Pe} \to +\infty$, it implies $\Gamma \to 0^+$ since infinite values for the density, velocity or characteristic length make no physical sense. Therefore the difussion coefficient tends to 0, which means the Laplacian term, linked to the diffusion process, is negligible. Dividing the PDE from (4.4) by Péclet's number results in the following equation

$$\phi_x + \phi_y = 0 \quad \text{in } \Omega \tag{4.6}$$

The following natural step would be considering equation (4.6) with g as boundary condition on all $\partial\Omega$, that is to say, the following problem:

$$\begin{cases} \phi_x + \phi_y = 0 & \text{in } \Omega \\ \phi = g & \text{on } \partial\Omega \end{cases} \tag{4.7}$$

Nonetheless, problem (4.7) is "overdetermined", which means a part of the boundary condition is unnecessary due to the geometric properties of the PDE as we shall see. In order to obtain a problem we can solve, take the curve $C = ([0, L] \times \{0\}) \cup (\{0\} \times (0, L])$, and let $\tilde{g} \colon C \subset \mathbb{R}^2 \to \mathbb{R}$ be

$$\tilde{g}(x, y) = \begin{cases} \phi_{\text{low}} & \text{if } (x, y) \in [0, L] \times \{0\} \\ \phi_{\text{high}} & \text{if } (x, y) \in \{0\} \times (0, L] \end{cases} \tag{4.8}$$

which is the restriction of g to C, that is to say, $\tilde{g} = g|_C$. The resulting Cauchy problem is

$$\begin{cases} \phi_x + \phi_y = 0 & \text{in } \Omega \\ \phi = \tilde{g} & \text{on } C \end{cases} \tag{4.9}$$

The PDE from (4.9) is known as the transport equation, which is a first–order linear PDE. In our case it has constant coefficients, making it easier to solve analitically.[1]

Definition 4.1. A classical solution to problem (4.9) is a function $\phi \colon \overline{\Omega} \to \mathbb{R}$ such that:

(i) $\phi \in \mathcal{C}^1(\overline{\Omega})$, i.e. ϕ is a $\mathcal{C}^1(\Omega)$ function that admits a $\mathcal{C}^1$ extension to an open neighbourhood of every point in $\partial\Omega$,

(ii) ϕ satisfies the PDE and

[1] Actually equation (4.9) is not the transport equation, since neither x nor y are time variables but space variables. The homogeneous transport equation is $u_t + b \cdot \nabla u = 0$ in $\mathbb{R}^n \times (0, \infty)$ where $u \colon \mathbb{R}^n \times [0, \infty) \to \mathbb{R}$ is the unknown. The importance of the time variable is that it gives the "cylinder" $\mathbb{R}^n \times [0, \infty) \subset \mathbb{R}^{n+1}$ where the equation occurs. Nonetheless problem (4.9) takes place in the domain $(0, L) \times (0, L)$ which is a cylinder as well, hence the y variable can be regarded as a time variable and solve the equation.

(iii) ϕ satisfies the boundary conditions.

In order to find the solution to (4.9), we will assume ϕ is a $\mathcal{C}^1(\overline{\Omega})$ function. Once we find the solution we will be able to tell whether ϕ is a classical solution. Moreover, so as to find a candidate of solution, we will make some assumptions motivated by intuiton and with lack of rigour, and later we shall justify them properly. This is a common practice in PDE theory.

We introduce some notation that will be useful. Given m vectors $\mathbf{w}_1, \dots, \mathbf{w}_m \in \mathbb{R}^n$, the set $[\mathbf{w}_1, \dots, \mathbf{w}_m] = \{\sum_{i=1}^m \lambda_i \mathbf{w}_i \mid \lambda_1, \dots, \lambda_m \in \mathbb{R}\}$ is the vector subspace of $\mathbb{R}^n$ spanned by $\mathbf{w}_1, \dots, \mathbf{w}_m$. If $W \subset \mathbb{R}^m$ is a vector subspace, $W^\perp = \{v \in \mathbb{R}^n \mid v \cdot w = 0 \ \forall w \in W\}$ is the vector subspace orthogonal to W.

The method we will follow to find the solution is known as the method of characteristics. Using $\boldsymbol{\nabla}\phi$ we can write the PDE as

$$(1,1) \cdot \boldsymbol{\nabla}\phi(x,y) = (1,1) \cdot \begin{pmatrix} \phi_x(x,y) \\ \phi_y(x,y) \end{pmatrix} = \phi_x(x,y) + \phi_y(x,y) = 0 \tag{4.10}$$

Recall from vector calculus that the gradient vector of ϕ gives the direction of maximum growth of ϕ at each point, whilst a non–zero vector $\mathbf{w} \in [\boldsymbol{\nabla}\phi(x,y)]^\perp$ provides the direction at (x,y) along which ϕ remains constant. Equation (4.10) tells us that at each point $(x,y) \in \Omega$, ϕ is constant along the direction given by $(1,1)$. To prove this, we exploit the fact that the PDE is first–order linear and use the chain rule to rewrite (4.10). Consider a $\mathcal{C}^1$ mapping $\alpha(s) = (\alpha_1(s), \alpha_2(s))$ such that $\alpha_1' = \alpha_2' = 1$ for all s. Since α is a mapping from some subset of $\mathbb{R}$ to $\mathbb{R}^2$, its image

$$A = \operatorname{Im}\alpha = \left\{(x,y) \in \mathbb{R}^2 \mid x = \alpha_1(s),\ y = \alpha_2(s),\ s \in \mathbb{R}\right\} \tag{4.11}$$

can be thought of as a $\mathcal{C}^1$ curve. Moreover, we may choose the curve to pass through $\Omega \cup C$ as we shall see in a moment. The restriction of ϕ to A, given by the composition $\varphi = \phi \circ \alpha \colon I \subset \mathbb{R} \to \mathbb{R}$, is also a $\mathcal{C}^1$ function as it is composition of $\mathcal{C}^1$ functions. By the chain rule,

$$\frac{\mathrm{d}}{\mathrm{d}s}\varphi(s) = \frac{\mathrm{d}}{\mathrm{d}s}\phi(\alpha_1(s), \alpha_2(s)) = \phi_x(\alpha_1(s), \alpha_2(s))\alpha_1'(s) + \phi_y(\alpha_1(s), \alpha_2(s))\alpha_2'(s) = \phi_x + \phi_y = 0 \tag{4.12}$$

where the last equality holds whenever $\alpha(s) \in \Omega$. Equation (4.12) implies that ϕ is constant on every connected component of $A \cap \Omega$, thereby proving our claim.

The following step is to find the curve A. Consider the mapping

$$\begin{aligned} f \colon \mathbb{R}^3 &\longrightarrow \mathbb{R}^2 \\ (s,x,y) &\longmapsto f(s,x,y) = (1,1) \end{aligned} \tag{4.13}$$

By taking a point $(x_0, y_0) \in \Omega \cup C$, we can find the curve A passing by (x_0, y_0):

$$\begin{cases} \alpha' = f(s,\alpha) = (1,1) & \text{in } I \subset \mathbb{R} \\ \alpha(0) = (x_0, y_0) \end{cases} \tag{4.14}$$

The function f is constant, therefore is Lipschitz continuous on (x,y) and uniformly with respect to s, thus the solution to (4.14) exists and is unique due to the Picard–Lindelöf Theorem (Theorem C.5). In addition, it is given by

$$\alpha(s) = (x_0 + s, y_0 + s) = (x_0, y_0) + s(1,1) \quad s \in I \subset \mathbb{R} \tag{4.15}$$

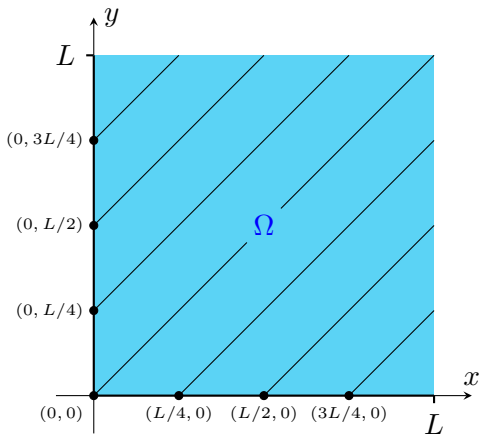


Figure 4.2. Some of the lines given by (4.15) with initial condition $(x_0, y_0) \in C$ extended to the top and right boundaries of Ω.

whence A is the line passing by (x_0, y_0) with director subspace $[(1, 1)]$. Moreover A is not a single line, but rather a family of lines with different initial condition. Hereinafter, we will take $(x_0, y_0) \in C$[2]. To distinguish the solutions (4.14) we will denote them by $\alpha(s; x_0, y_0)$, and the curves by $A_{(x_0,y_0)}$.

We claim that a solution (4.15) can be extended so that $\alpha(s; x_0, y_0)$ eventually reaches the top or right boundaries of Ω as shown in figure 4.2. Take $T > 0$ and $\delta > 0$ to be some constants to be determined and let $V = [0, T] \times \overline{B((x_0, y_0), \delta)} \subset \mathbb{R}^3$. Since f is a constant function, we have

$$M = \sup_{(s,x,y)\in V} \|f(s, x, y)\| = \max_{(s,x,y)\in V} \|f(s, x, y)\| = \sqrt{2} \tag{4.16}$$

Again, by the Picard–Lindelöf theorem, the solution (4.15) exists for $s \in I = [0, T_0] \subset \mathbb{R}$ and remains in $\overline{B((x_0, y_0), \delta)}$ where $T_0 = \min\left\{T, \frac{\delta}{M}\right\}$. By taking $\delta = \sqrt{2}L$ and $T = L$, applying the theorem we obtain $T_0 = L$ and the solution stays in $\overline{B((x_0, y_0), \sqrt{2}L)}$. Since $\sqrt{2}L$ is the maximum of the distances between two points belonging to $\overline{\Omega}$, we have proved our claim. As a consequence, by changing (x_0, y_0) we can fill $\overline{\Omega}$ with these curves. All except one of the solutions $\alpha(s; x_0, y_0)$ actually exit $\overline{\Omega}$, however we do not care about the part of the curve outside $\overline{\Omega}$[3].

Up to now, we have found out the following:

(i) The lines given by $\alpha(s; x_0, y_0)$ can be extended so that both ends touch $\partial\Omega$. The implicit form of these is

$$A_{(x_0,y_0)} \colon x - y = x_0 - y_0, \quad (x_0, y_0) \in C \tag{4.17}$$

(ii) The lines (4.17) fill $\overline{\Omega}$.

(iii) By equation (4.12), the function ϕ is constant on every line (4.17).

The curves (4.17) are known as the characteristic lines (or simply characteristics) of problem (4.9). Some of them are picture in figure 4.3.

We know the value of ϕ at $(x_0, y_0) \in C$ and ϕ is constant along the curve $A_{(x_0,y_0)}$. Therefore the value of ϕ at $(x, y) \in A_{(x_0,y_0)}$ is $\phi(x, y) = \phi(x_0, y_0) = \tilde{g}(x_0, y_0)$. As $(x_0, y_0) \in C$ implies either $x_0 = 0$ or $y_0 = 0$ (or both), we have the following:

[2] At the points $(x_0, y_0) \in C$ we have the boundary condition, i.e. we have information about the solution ϕ.
[3] We have such freedom to choose the constants T and δ because f is a constant function.

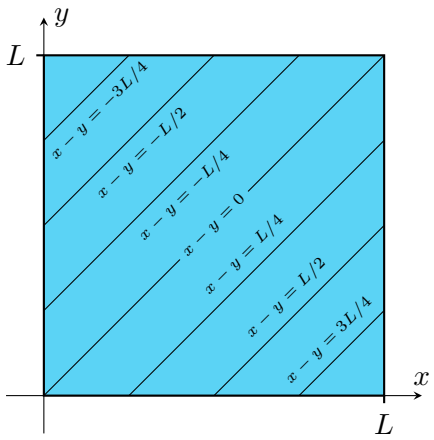


Figure 4.3. Some characteristics of problem (4.9).

- If $y \leq x$ then $\phi(x, y) = \phi(x - y, 0) = \tilde{g}(x - y, 0)$.
- If $y > x$ then $\phi(x, y) = \phi(0, y - x) = \tilde{g}(0, y - x)$.

With this in mind, the solution to (4.9) is:

$$\phi(x, y) = \begin{cases} \tilde{g}(x - y, 0) = \phi_{\text{low}} & \text{if } y \leq x \\ \tilde{g}(0, y - x) = \phi_{\text{high}} & \text{if } y > x \end{cases} \quad (x, y) \in \overline{\Omega} \tag{4.18}$$

Intuitively, the characteristics give the paths in $\mathbb{R}^2$ through which the information of the boundary conditions is transported.

After finding the solution, we should check if $\phi \in \mathcal{C}^1(\overline{\Omega})$. First consider the case when $\phi_{\text{low}} = \phi_{\text{high}}$.

Theorem 4.2. Assume $\phi_{\text{low}} = \phi_{\text{high}}$. Then the solution to problem (4.9) exists and is unique. Moreover it is a solution in the classical sense.

Proof. We have proved the existence of a solution by giving formula (4.18). The uniqueness is a consequence of the method of characteristics. In it we have seen that ϕ is constant on each the characteristic, then we have found the equation of characteristics and and proved that given an initial condition $(x_0, y_0) \in C$, the curve is unique. Finally ϕ is a $\mathcal{C}^1(\Omega) \cap \mathcal{C}(\overline{\Omega})$ function because it is constant on $\overline{\Omega}$ and clearly satisfies the boundary condition by construction and the PDE. □

Assume $\phi_{\text{low}} < \phi_{\text{high}}$, then ϕ is not continuous on the segment $\{x - y = 0\} \cap \overline{\Omega}$ thus it cannot be a differentiable function, implying that function (4.18) is not a classical solution. This could be warned from the beginning, since any function satisfying the boundary condition of problem (4.9) is not continuous at $(0, 0)$.

4.2.2 Weak solution for $\text{Pe} = \infty$

As we have seen in the previous subsection, problem

Definition 4.3. A function $\psi \colon \overline{\Omega} \to \mathbb{R}$ is said to be a weak solution of problem (4.9) if for all test functions $\psi \in \mathcal{C}_c^1(\overline{\Omega})$ the following integral equation is satisfied:

$$\int_{\Omega} \tag{4.19}$$

A function $\psi\colon \overline{\Omega} \to \mathbb{R}$ is said a weak solution of (4.9) if

$$\int_{\Omega}$$

Notice that each characteristic starting on C_1 ends on C_1, and the same holds for C_2.

y definition of the Cauchy problem, ϕ is constant on C_1 and on C_2. Therefore the value of ϕ on the characteristic $x - y = c$ is the value that g takes on the part of the boundary the characteristic intersects.

4.2.3 Classical solution for $0 \leq \mathrm{Pe} < \infty$

Now we focus in problem (4.4) with $0 \leq \mathrm{Pe} < \infty$ and $\phi_{\text{low}} < \phi_{\text{high}}$. The PDE we are dealing with is a second–order elliptic equation with constant coefficients. First of all, we would like to know whether a classical solution exists:

Definition 4.4. A classical solution to problem (4.9) is a function $\phi\colon \overline{\Omega} \to \mathbb{R}$ such that:

(i) $\phi \in \mathcal{C}^2(\overline{\Omega})$, i.e. ϕ is a $\mathcal{C}^2(\Omega)$ function that admits a $\mathcal{C}^2$ extension to an open neighbourhood of every point in $\partial\Omega$,

(ii) ϕ satisfies the PDE and

(iii) ϕ satisfies the boundary conditions.

The function g giving the boundary condition is not continuous at $(0,0)$ nor at (L,L) unless $\phi_{\text{low}} = \phi_{\text{high}}$. Therefore problem (4.4) cannot have a classical solution, however it might have a solution in the weak sense.

4.2.4 Weak solution for $0 \leq \mathrm{Pe} < \infty$

The theory that studies the existence and uniqueness of weak solutions to Cauchy problems involving second–order elliptic equations requires a prior knowledge in measure theory and functional analysis. The interested reader can consult appendix A for a quick reference in some basic concepts of measure theory. We will not introduce functional analysis concepts in order not to complicate the exposition needlessly. Rather than focusing on problem (4.4), we shall study the theory for a general second–order elliptic equation and then particularize to our case. The reference for this subsection is Lawrence C. Evans' excellent book on Partial Differential Equations [5], in particular chapter 6.

General theory for weak solutions

Consider the Cauchy problem

$$\begin{cases} Lu = f & \text{in } U \\ u = 0 & \text{on } \partial U \end{cases} \tag{4.20}$$

where $U \subset \mathbb{R}^n$ is an open bounded subset, $u\colon \overline{U} \to \mathbb{R}$ is the unknown function (shall not be confused with the x–component of the velocity field) and $f\colon U \to \mathbb{R}$ is a given function. L is a second–order

partial differential operator having one of the two following forms

$$Lu = -\sum_{i,j=1}^{n} (a^{ij}(x)u_{x_i})_{x_j} + \sum_{i=1}^{n} b^i(x)u_{x_i} + c(x)u \tag{4.21}$$

$$Lu = -\sum_{i,j=1}^{n} a^{ij}(x)u_{x_ix_j} + \sum_{i=1}^{n} b^i(x)u_{x_i} + c(x)u \tag{4.22}$$

being a^{ij}, b^i, c for $i,j = 1,\dots,n$ given functions. Aditionally, we shall assume a^{ij}, b^i, $c \in L^\infty(U)$ for all $i,j = 1,\dots,n$ and $f \in L^2(U)$. We say L is in divergence form when is given by (4.21), whereas L is in non–divergence form when expressed by (4.22).

Proposition 4.5. Whenever $a^{ij} \in \mathcal{C}^1(U)$ for $i,j = 1,\dots,n$, a partial differential operator in divergence form can be rewritten in non–divergence form and viceversa.

Proof. Assume L is given in divergence form, then

$$\begin{aligned} Lu &= -\sum_{i,j=1}^{n} (a^{ij}(x)u_{x_i})x_j + \sum_{i=1}^{n} b^i(x)u_{x_i} + c(x)u \\ &= -\sum_{i,j=1}^{n} \left\{ a^{ij}(x)u_{x_ix_j} + a^{ij}_{x_j}(x)u_{x_i} \right\} + \sum_{i=1}^{n} b^i(x)u_{x_i} + c(x)u \\ &= -\sum_{i,j=1}^{n} a^{ij}(x)u_{x_ix_j} + \sum_{i=1}^{n} \left\{ b^i(x) - \sum_{j=1}^{n} a^{ij}_{x_j}(x) \right\} u_{x_i} + c(x)u \end{aligned}$$

which is in non–divergence form. Conversely, assume L is given in non–divergence form, thus

$$\begin{aligned} Lu &= -\sum_{i,j=1}^{n} a^{ij}(x)u_{x_ix_j} + \sum_{i=1}^{n} b^i(x)u_{x_i} + c(x)u \\ &= -\sum_{i,j=1}^{n} \left\{ a^{ij}(x)u_{x_ix_j} + a^{ij}_{x_j}(x)u_{x_i} - a^{ij}_{x_j}(x)u_{x_i} \right\} + \sum_{i=1}^{n} b^i(x)u_{x_i} + c(x)u \\ &= -\sum_{i,j=1}^{n} (a^{ij}(x)u_{x_i})_{x_j} + \sum_{i=1}^{n} \left\{ b^i(x) + \sum_{j=1}^{n} a^{ij}_{x_j}(x) \right\} u_{x_i} + c(x)u \end{aligned}$$

giving the divergence form of L. □

Definition 4.6.

(i) We say the partial differential operator L is symmetric provided

$$a^{ij} = a^{ji} \tag{4.23}$$

for all $i,j = 1,\dots,n$.

(ii) We say the partial differential operator L is (uniformly) elliptic if there exists a constant $\theta > 0$ such that

$$\sum_{i,j=1}^{n} a^{ij}(x)\,\xi_i\xi_j \geq \theta\|\xi\|^2 \tag{4.24}$$

for a.e. $x \in U$ and all $\xi \in \mathbb{R}^n$.

Hereinafter we will assume operator L satisfies both the symmetry and the uniform ellipticity conditions. Before giving the definition of weak solution, we shall gain some intuition about it. Suppose L is given in divergence form. Let us assume that u is a smooth solution to (4.20), i.e. u is a differentiable enough function. Let $v \in \mathcal{C}_c^\infty(U)$ be a smooth test function, where $\mathcal{C}_c^\infty(U)$ is the set of infinitely differentiable functions with compact support in U. We multiply $Lu = f$ by v and integrate over U:

$$\int_U -\sum_{i,j=1}^n (a^{ij}(x)u_{x_i})_{x_j} v \,\mathrm{d}x + \int_U \sum_{i=1}^n b^i(x) u_{x_i} v \,\mathrm{d}x + \int_U c(x)uv \,\mathrm{d}x = \int_U fv \,\mathrm{d}x \tag{4.25}$$

We may rewrite the first term on the left–hand side integrating by parts

$$\begin{aligned}\int_U -\sum_{i,j=1}^n (a^{ij}(x)u_{x_i})_{x_j} v \,\mathrm{d}x &= \int_U \sum_{i,j=1}^n a^{ij}(x) u_{x_i} v_{x_j} \,\mathrm{d}x - \int_{\partial U} \sum_{i,j=1}^n a^{ij}(x) u_{x_i} v \nu^j \,\mathrm{d}\sigma \\ &= \int_U \sum_{i,j=1}^n a^{ij}(x) u_{x_i} v_{x_j} \,\mathrm{d}x \end{aligned} \tag{4.26}$$

In (4.26), ν^j denotes the j–th component of the exterior normal vector to ∂U and $\mathrm{d}\sigma$ is the "surface" element of ∂U. The integral over ∂U is zero as v vanishes on it as a consequence of being in $\mathcal{C}_c^\infty(U)$. Introducing (4.26) in (4.25) yields

$$\int_U \sum_{i,j=1}^n a^{ij} u_{x_i} v_{x_j} + \sum_{i=1}^n b^i u_{x_i} v + cuv \,\mathrm{d}x = \int_U fv \,\mathrm{d}x \tag{4.27}$$

The left–hand side of (4.27) is the bilinear form

$$B[u,v] = \int_U \sum_{i,j=1}^n a^{ij} u_{x_i} v_{x_j} + \sum_{i=1}^n b^i u_{x_i} v + cuv \,\mathrm{d}x \tag{4.28}$$

for $u, v \in H_0^1(U)$, and is associated to the operator L in divergence form. Thus equality (4.27) is rewritten as

$$B[u,v] = (f,v) \tag{4.29}$$

where $(f,v) = \int_U fv \,\mathrm{d}x$ is the scalar product in $L^2(U)$. Once we have introduced $B[\cdot,\cdot]$, we can define the concept of weak solution.

Definition 4.7. A function $u \in H_0^1(U)$ is weak solution of the boundary value problem (4.20) if

$$B[u,v] = (f,v) \quad \text{for all } v \in H_0^1(U) \tag{4.30}$$

The intuition behind the definition of weak solution is the following: when a function $u \in H_0^1(U)$ is fixed, and equality (4.30) holds for any other function $v \in H_0^1(U)$, following the development we have made, necessarily u is a solution to $Lu = f$ in $U \subset \mathbb{R}^n$. The theorem that makes this intuition precise, although some further hypothesis on B are necessary, is the Lax–Milgram theorem.

Before stating the theorem on the existence and uniqueness of solution to (4.20), other results such as Lax–Milgram's theorem or energy estimates are necessary, however we shall not present them here so as not to complicate the exposition. The interested reader can refer to [5]. The following is Theorem 4 from section 6.2 of Evan's book.

Theorem 4.8 (Second Existence Theorem for weak solutions)**.**

(1) Precisely one of the following statements holds:

(1.1) For each $f \in L^2(U)$ there exists a unique weak solution u of the boundary value problem

$$\begin{cases} Lu = f & \text{in } U \\ u = 0 & \text{on } \partial U \end{cases} \tag{4.31}$$

(1.2) There exists a weak solution $u \not\equiv 0$ of the homogeneous problem

$$\begin{cases} Lu = 0 & \text{in } U \\ u = 0 & \text{on } \partial U \end{cases} \tag{4.32}$$

(2) Furthermore, should assertion (1.2) hold, the dimension of the subspace $N \subset H_0^1(U)$ of weak solutions of (4.32) is finite and equals the dimension of the subspace $N^* \subset H_0^1(U)$ of weak solutions of

$$\begin{cases} L^* v = 0 & \text{in } U \\ v = 0 & \text{on } \partial U \end{cases} \tag{4.33}$$

(3) Finally, the boundary–value problem (4.31) has a weak solution if and only if

$$(f, v) = 0 \quad \text{for all } v \in N^* \tag{4.34}$$

Theorem 4.9 (Infinite differentiability in the interior)**.** Assume

$$a^{ij},\, b^i,\, c \in \mathcal{C}^\infty(U) \quad (i, j = 1, \ldots, n)$$

and

$$f \in \mathcal{C}^\infty(U)$$

Suppose $u \in H^1(U)$ is a weak solution of the elliptic PDE

$$Lu = f \quad \text{in } U$$

Then

$$u \in \mathcal{C}^\infty(U)$$

Weak solution for $0 \leq \mathrm{Pe} < \infty$

Now we shall apply the previous theory to problem (4.4). Observe that the PDE is equivalent to

$$-\Delta\phi + (\phi_x + \phi_y)\beta\,\mathrm{Pe} = -\phi_{xx} - \phi_{yy} + (\phi_x + \phi_y)\beta\,\mathrm{Pe} = 0 \tag{4.35}$$

So as to express (4.35) with an operator L such as the ones described above, we define

$$\begin{aligned} a^{11} = 1 \quad a^{12} = 0 \quad b^1 = \beta\,\mathrm{Pe} \quad c = 0 \\ a^{21} = 0 \quad a^{22} = 1 \quad b^2 = \beta\,\mathrm{Pe} \end{aligned} \tag{4.36}$$

therefore the operator L encoding (4.35) is simply

$$L\phi = -\sum_{i,j=1}^{2} (a^{ij}\phi_{x_i})_{x_j} + \sum_{i=1}^{n} b^i \phi_{x_i} + c\phi = -\phi_{xx} - \phi_{yy} + (\phi_x + \phi_y)\beta\,\mathrm{Pe} \tag{4.37}$$

where $x_1 \equiv x$, $x_2 \equiv y$. Moreover, since the functions in (4.36) are constant, these belong to $L^\infty(\Omega)$ and the divergence form and non–divergence form of (4.37) are the same. It is obvious that L is symmetric, as $a^{12} = a^{21} = 0$, and it is uniformly elliptic with $\theta = 1$ as a result of

$$\sum_{i,j=1}^{2} a^{ij}\xi_i\xi_j = {\xi_1}^2 + {\xi_2}^2 = \theta\|\xi\|^2 \tag{4.38}$$

for all $\xi \in \mathbb{R}^2$.

Now we shall transform the boundary–value problem (4.4) into a problem having the form (4.20). Observe that the boundary condition in (4.20) is homogeneous, i.e. $u = 0$ on ∂U, whilst in (4.4) this is not true. Assume ϕ is a weak solution to (4.4), then $\tilde{\phi} = \phi - g$ is a solution to

$$\begin{cases} L\tilde{\phi} = 0 & \text{in } \Omega \\ \tilde{\phi} = 0 & \text{on } \partial\Omega \end{cases} \tag{4.39}$$

since $L\tilde{\phi} = L\phi - Lg = L\phi$.

Remark. Notice that weak solutions given by theorem (4.8) are functions

By theorem 4.8, problem (4.39) has a weak solution. Nonetheless, we cannot specify whether there is unique solution or not due to dichotomy (1.1)–(1.2), which is known as the Fredholm alternative.

4.3 Numerical solution

In this section we present the numerical solution of problem (4.4) for several Péclet numbers. The width and height of the domain are $L = 1$ m and the velocity of the flow is $u = 1$ m/s. The density is kept constant at $\rho = 1000$ kg/m^3, therefore Péclet's number is varied by chaging the diffusion coefficient Γ. The boundary conditions are $\phi_{\text{low}} = 0$ and $\phi_{\text{high}} = 1$. A uniform mesh of $N = 200$ nodes has been used to discretize the domain, with a tolerance of 10^{-12} as a stop criterion for the Gauss–Seidel algorithm. The Upwind–Difference Scheme (UDS) has been chosen to compute the convective properties.

Figure 4.4 shows the solution to the diagonal case problem for Pe = 1. Transport and diffusion have a similar strength as can be seen in the central zone of the domain Ω. There is clearly a transport phenomena carrying the fluid from the lower left corner to the upper right corner of Ω, but there is also mixing due to the diffusion process. Note that the solution is not continuous at the lower left and upper right corners because of the sudden jump from ϕ_{low} to ϕ_{high}. The zone around the upper left corner does not seem affected by the diffusion process as it is far from the boundary where $\phi = \phi_{\text{low}}$. The same applies to the zone close to the lower right corner.

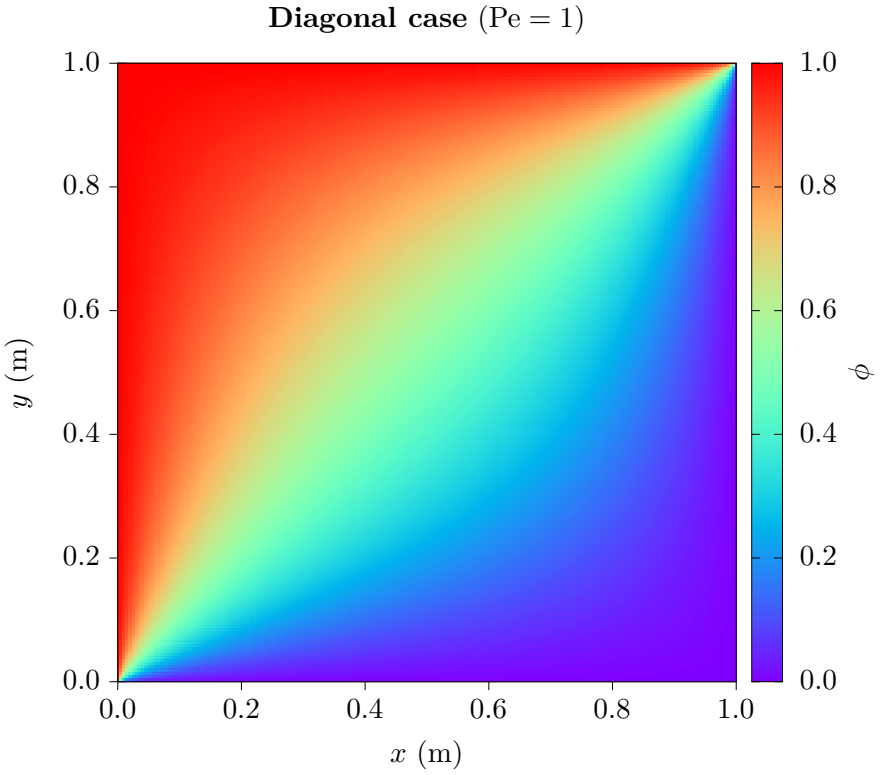
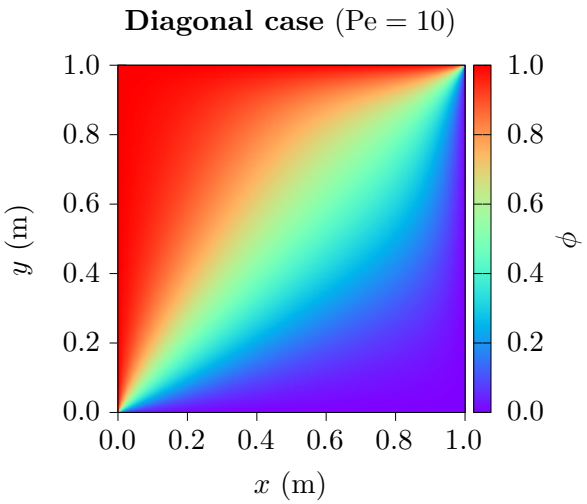
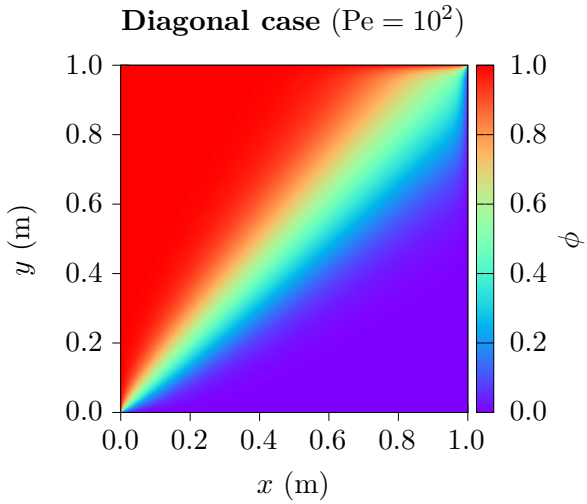


Figure 4.4. Numerical solution to the diagonal case for Pe = 1.

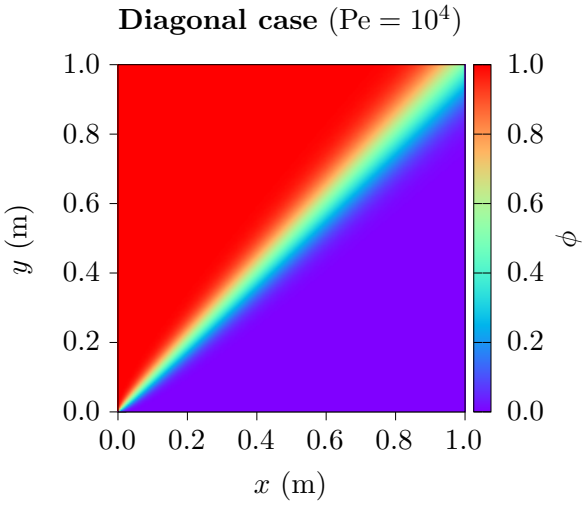
Figures (4.5a) to (4.5c) show the diagonal case solution for Pe = 10, 10^2, 10^4 and 10^9. The solution for Pe = 10 (figure 4.5a) has a similar appearance to the solution for Pe = 1 (figure 4.4). As Péclet's number grows, the transport process takes over the diffusion process. Therefore the diffusion zone, which is centered in the diagonal along the fluid flow, tends to shrink. This change in the behaviour of the solution can be observed by comparing the cases for Pe = 10 (figure 4.5a) and Pe = 10^2 (figure 4.5b). For Pe = 10^4 the diffusion zone becomes even narrower. Beyond Pe = 10^4 there are no obvious changes in the solution, as can be checked by looking at the case for Pe = 10^9 (figure 4.5d).



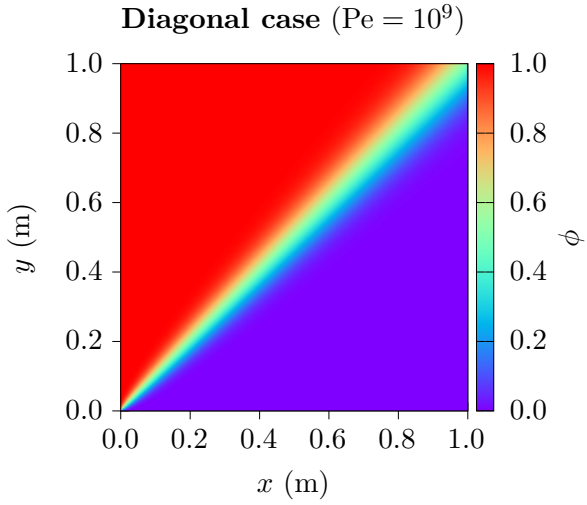
(a) Numerical solution to the diagonal case for Pe = 10.



(b) Numerical solution to the diagonal case for Pe = 10^2.



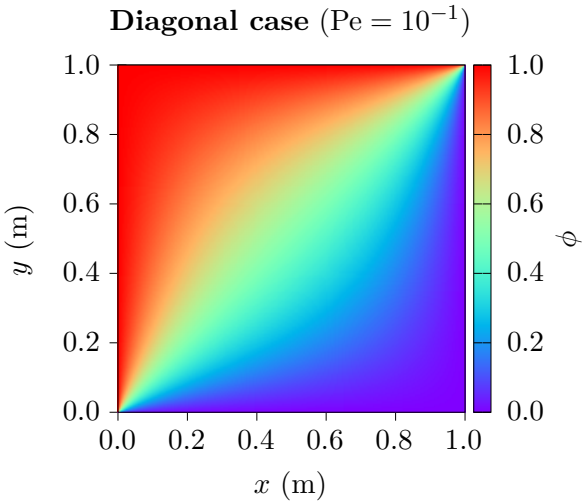
(c) Numerical solution to the diagonal case for Pe = 10^3.



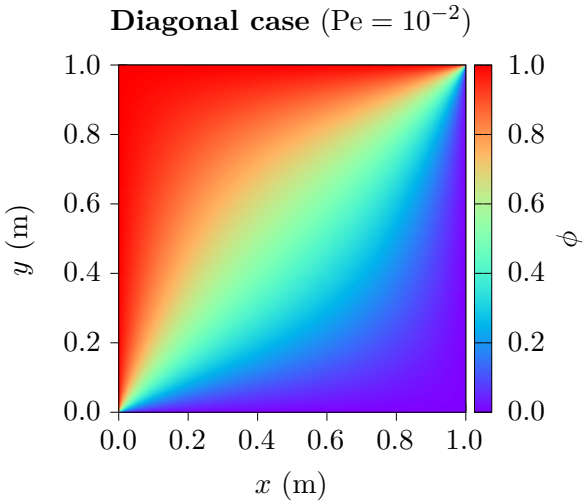
(d) Numerical solution to the diagonal case for Pe = 1^90.

Figure 4.5. Numerical solution to the diagonal case for Pe = 10, $10^2, 10^4$ and 10^9.

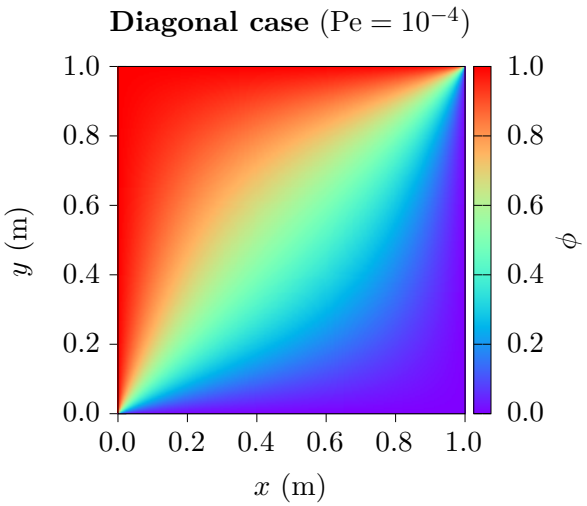
Figures 4.6a to 4.6d show the diagonal case solution to Pe $= 10^{-1}$, 10^{-2}, 10^{-4} and 10^{-9}. As it can be observed, all the solutions have a similar appearance to that for Pe $= 1$ (figure 4.4), whence it can be deduced that reducing Péclet's number has not obvious effect.



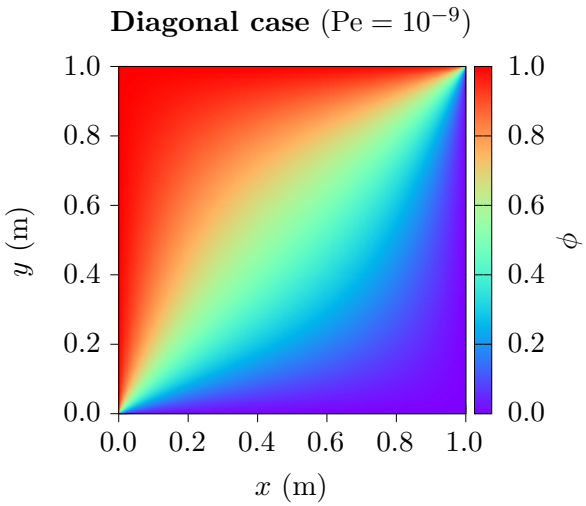
(a) Numerical solution to the diagonal case for Pe $= 10^{-1}$.



(b) Numerical solution to the diagonal case for Pe $= 10^{-2}$.



(c) Numerical solution to the diagonal case for Pe $= 10^{-4}$.



(d) Numerical solution to the diagonal case for Pe $= 10^{-9}$.

Figure 4.6. Numerical solution to the diagonal case for Pe $= 10^{-1}$, 10^{-2}, 10^{-4} and 10^{-9}.

5 Smith–Hutton case

5.1 Statement

This section deals with the steady state version of the problem proposed by Smith and Hutton (1982) described in [12]. The problem takes place in the domain $\Omega = (-L, L) \times (0, L) \subset \mathbb{R}^2$ where $L > 0$ is a constant length. Both density and diffusion coefficient are assumed to be constant and known values. In Ω the steady state version of the general convection–diffusion equation with no source term is considered, that is,

$$\frac{\rho}{\Gamma}\mathbf{v} \cdot \boldsymbol{\nabla}\phi = \Delta\phi \tag{5.1}$$

On the boundary of Ω the following conditions are prescribed:

- $\phi = 1 + \tanh(10(2x + 1))$ on $C_1 = [-L, 0] \times \{0\}$ (inlet flow).
- $\phi = 1 - \tanh(10)$ on $C_2 = (\{-L\} \times (0, L)) \cup ([-L, L] \times \{L\}) \cup (\{L\} \times [0, L))$.
- $\dfrac{\partial\phi}{\partial y} = 0$ on $C_3 = (0, L) \times \{0\}$ (outlet flow).

Notice that the curves C_1, C_2, C_3 give a partition of $\partial\Omega$. To encode the first two boundary conditions in a compact manner, we define the function $g\colon C_1 \cup C_2 \to \mathbb{R}$ by

$$g(x, y) = \begin{cases} 1 + \tanh(10(2x + 1)) & \text{if } (x, y) \in C_1 \\ 1 - \tanh(10) & \text{if } (x, y) \in C_2 \end{cases} \tag{5.2}$$

The velocity field is given by $u = 2y(1 - x^2)$ and $v = -2x(1 - y^2)$. The Cauchy problem resulting from the PDE (5.1) and the boundary conditions is given by (5.3) and is summarized in figure 5.1.

$$\begin{cases} \Delta\phi - \dfrac{\rho}{\Gamma}\mathbf{v} \cdot \boldsymbol{\nabla}\phi = 0 & \text{in } \Omega \\ \phi = g & \text{on } C_1 \cup C_2 \\ \dfrac{\partial\phi}{\partial y} = 0 & \text{on } C_3 \end{cases} \tag{5.3}$$

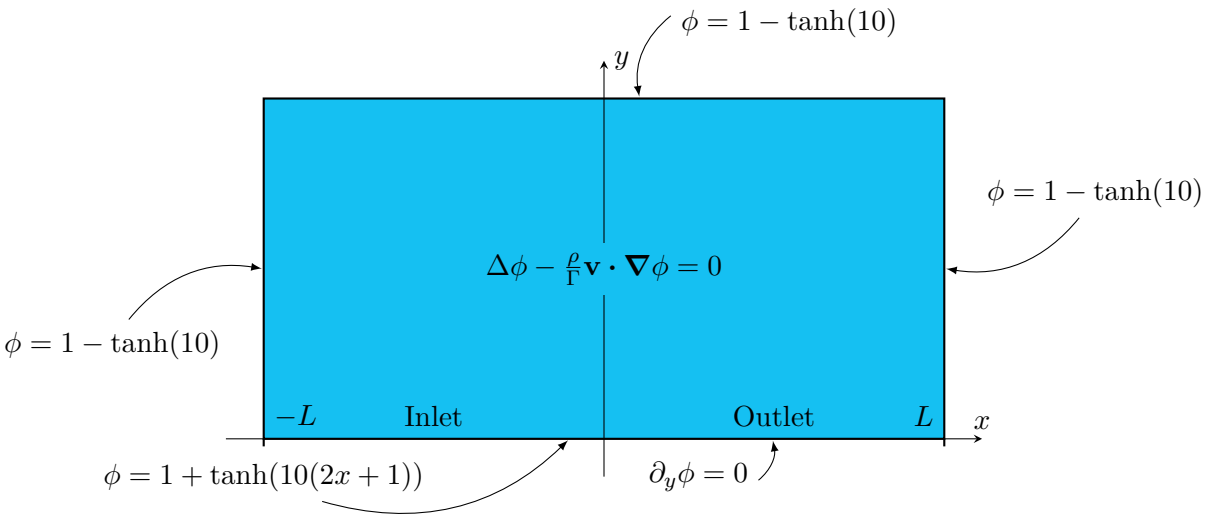


Figure 5.1. Cauchy problem for the diagonal flow case.

5.2 Velocity field

The velocity field for the Smith-Hutton case is given by $\mathbf{v} = 2y(1-x^2)\mathbf{i} - 2x(1-y^2)\mathbf{j}$. It verifies the incompressibility condition since it is divergence–free, i.e. $\boldsymbol{\nabla}\cdot\mathbf{v} = 0$. The only points where $\mathbf{v}$ vanishes are $(0,0)$ and $(\pm 1, \pm 1)$. In the case of $L < 1$, only $(0,0)$ belongs to $\overline{\Omega}$. Otherwise, $(0,0)$, $(-1,1)$ and $(1,1)$ the first three points belong to $\overline{\Omega}$.

Recall that the stream function is a mapping $\psi\colon U \subset \mathbb{R}^2 \to \mathbb{R}$, where U is some open subset of $\mathbb{R}^2$ containing Ω, such that

$$u = \frac{\partial \psi}{\partial y}, \quad v = -\frac{\partial \psi}{\partial x} \tag{5.4}$$

By choosing u and v to be the components of the velocity field $\mathbf{v}$ and integrating, one finds the function $\psi(x,y) = x^2 + y^2(1-x^2) + C$ where $C \in \mathbb{R}$ is a constant. Since the constant vanishes when differentiating, we may choose $C = -1$ so that we obtain the more nice looking

$$\psi(x,y) = -(1-x^2)(1-y^2) \tag{5.5}$$

In this and in the forthcoming sections we shall assume $L = 1$.

In order to find the analytical solution to problem (5.3) when $\Gamma = 0$, it will be useful to know how the streamlines are. Recall that the streamlines are defined as the curves tangent to the vector field $\mathbf{v}$ at each point. If $\alpha\colon I \subset \mathbb{R} \to \Omega$, $s \mapsto \alpha(s) = (x(s), y(s))$ is the parametrization of a curve in Ω, then it is a streamline provided it satisifes the following system of ODEs:

$$\begin{cases} x' = 2y(1-x^2) \\ y' = -2x(1-y^2) \end{cases} \tag{5.6}$$

Equivalently, the streamlines are the curves with normal vector $\boldsymbol{\nabla}\psi$ at every point. In order to find the streamlines, we can specify an initial condition and then pose an initial value problem. Consider the mapping $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ defined by

$$f(x,y) = \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix} = \begin{pmatrix} 2y(1-x^2) \\ -2x(1-y^2) \end{pmatrix} \tag{5.7}$$

Then the initial value problem for a streamline with initial condition $(x_0, y_0) \in \overline{\Omega}$ is

$$\begin{cases} \alpha' = f(\alpha(s)) \\ \alpha(0) = (x_0, y_0) \end{cases} \tag{5.8}$$

Proposition 5.1. The solution to the initial value problem (5.8) exists and is unique for every $(x_0, y_0) \in \overline{\Omega}$.

Proof. Let $U = B(0, 2L)$, thus $\overline{\Omega} \subset U$. Since both u and v are $\mathcal{C}^\infty(\mathbb{R}^2)$ functions, f is $\mathcal{C}^\infty(\mathbb{R}^2)$. The restriction of f to $\overline{U}$ is, in particular, a $\mathcal{C}^1(\overline{U})$ mapping. By theorem C.4, f is Lipschitz on $\overline{U}$. Now by the Picard–Lindelöf theorem (Theorem C.5) we deduce the existence and uniqueness of solution to (5.8). □

Finding the explicit solution to (5.8) might be difficult as the system is non–linear. Nonetheless we can solve it numerically so as to find the appearance of the streamlines. This is precisely what has been done to produce figure (5.2). The RK4 algorithm was applied to the IVP (5.8) for $L = 1$ m and initial conditions $x_0 = 0.10,\ 0.20,\ 0.30,\ 0.40,\ 0.50,\ 0.60,\ 0.70,\ 0.80,\ 0.90,\ 0.99$ m and $y_0 = 0$.

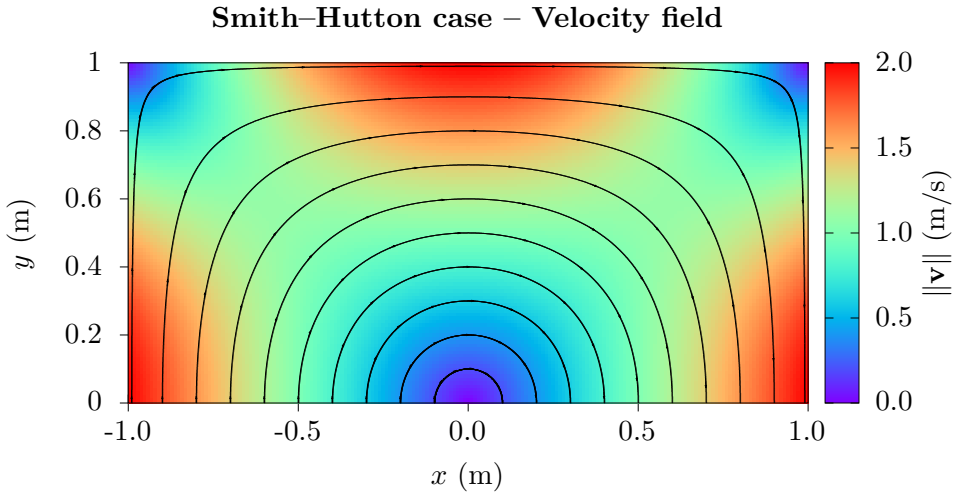


Figure 5.2. Norm of the Smith–Hutton velocity field and streamlines for $x_0 = 0.10,\ 0.20,\ 0.30,\ 0.40,\ 0.50,\ 0.60,\ 0.70,\ 0.80,\ 0.90$ and 0.99 m. The vectors tangent to the streamlines are normalized and then scaled down by a factor of $\sqrt{2}/50$.

5.3 Analytical solution

Unlike the diagonal case, in the Smith–Hutton problem there is no clear choice for a fluid velocity to compute Péclet's number. The average velocity of the flow field given by

$$\overline{v} = \frac{\iint_\Omega \|\mathbf{v}\| \,\mathrm{d}x\,\mathrm{d}y}{\iint_\Omega 1 \,\mathrm{d}x\,\mathrm{d}y} = \frac{1}{2L^2} \iint_\Omega \sqrt{4y^2(1-x^2)^2 + 4x^2(1-y^2)^2}\,\mathrm{d}x\,\mathrm{d}y \tag{5.9}$$

could be chosen as a representative velocity. However, it is unclear whether or not it is relevant for the behaviour of the solution. In view of the PDE from (5.3), it seems that the solution shall be governed by the quotient ρ/Γ.

5.3.1 General problem

5.4 Numerical solution

In this section we present the numerical solution to problem (5.3) for several values of the quotient ρ/Γ. The characteristic length taken is $L = 1$ m. The density is kept constant at $\rho = 1000$ kg/m^3 and the diffusion coefficient Γ is varied. A uniform mesh has been used to discretize the domain, with $N_x = 201$ nodes in the x–axis and $N_y = 101$ nodes in the y–axis. The tolerance to stop Gauss–Seidel's iteration has been of 10^{-12}. The convective properties have been evaluated applying the Power law Scheme.

As it has been said, the characteristic length of the problem is $L = 1$ m, while the characteristic velocity is unknown. Nonetheless, it must be constant as the velocity field $\mathbf{v}$ does not depend on time, hence Péclet's number depends on the quotient ρ/Γ. This implies that ρ/Γ gives an idea of the relation convection transport rate/diffusion transport rate.

Figure 5.3 shows the numerical solution to the Smith–Hutton case for $\rho/\Gamma = 1$. Both processes, transport and diffusion, apparently have a similar strength. There is clearly transport phenomena taking the information about ϕ from the inlet zone $(-1, 0] \times \{0\}$ to the outlet zone $(0, 1) \times \{0\}$. For instance, the inlet zone with $\phi \approx 1.0$ (green zone) occupies a rather small part of the inlet around $x = -0.5$ m. However, as the transport occurs the band corresponding to $\phi \approx 1$ becomes wider due to the diffusion process. This influence of diffusion can also be ssen for $\phi \approx 0.5$ (light blue band) and for $\phi \approx 1.5$ (orange band).

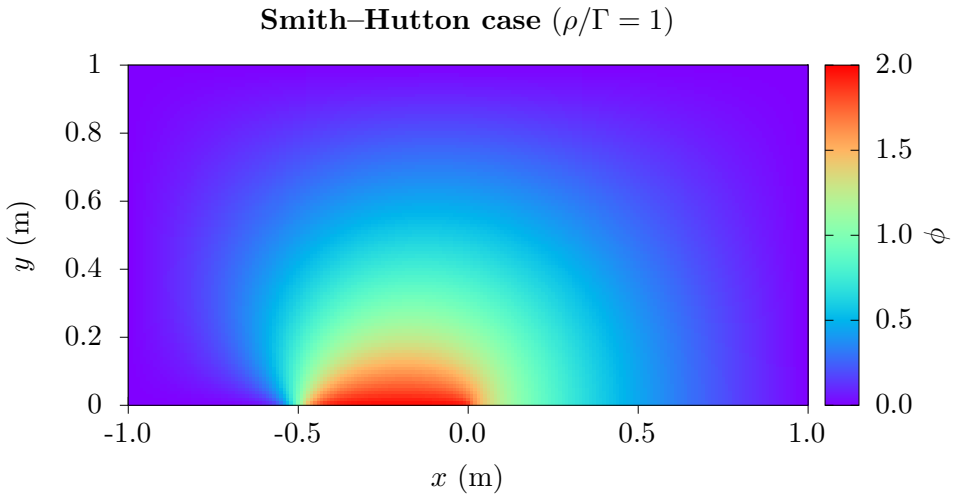


Figure 5.3. Numerical solution the the Smith–Hutton case for $\rho/\Gamma = 1$.

Figures 5.4 and 5.5 show the numerical solution to the Smith–Hutton problem for $\rho/\Gamma = 10$ and $\rho/\Gamma = 100$, respectively. Some differences between the solutions for $\rho/\Gamma = 1$ and $\rho/\Gamma = 10$, although not too obvious, can be spotted. Clearly the light blue, green and orange bands now occupy a larger portion of Ω, while the smooth transitions between them are still present. In contrast to the $\rho/\Gamma = 10$ case, the change for $\rho/\Gamma = 100$ is more apparent. Now the transition zones between the different color bands are thinner, implying the diffusion process has lost strength with respect to the transport process. However there is still diffusion, since the green band widens along the streamlines.

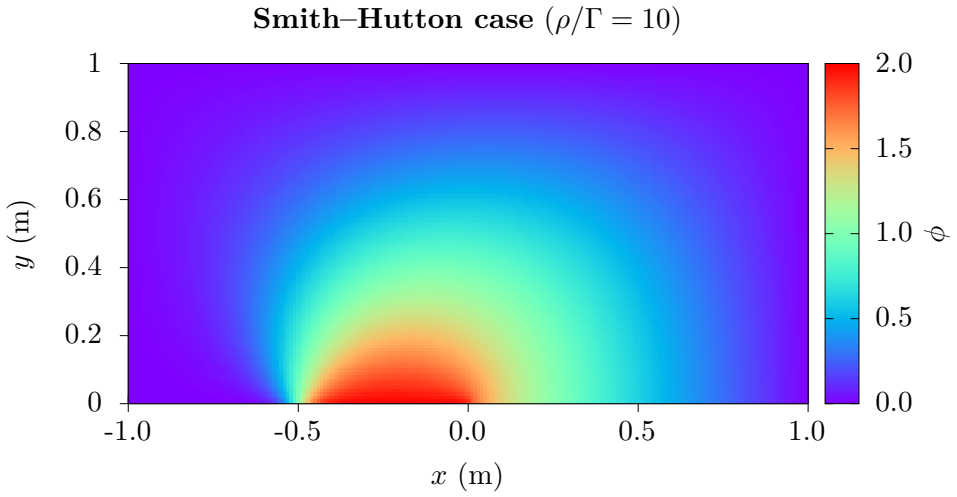


Figure 5.4. Numerical solution the the Smith–Hutton case for $\rho/\Gamma = 10$.

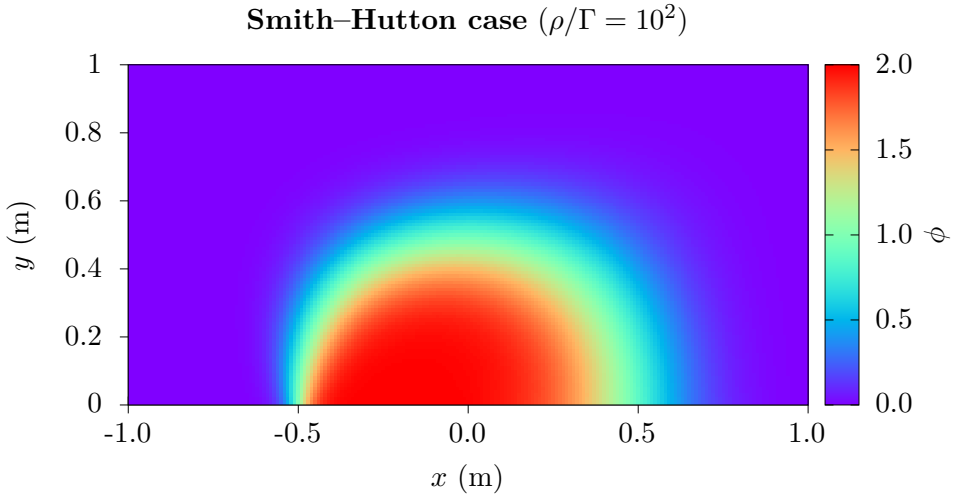


Figure 5.5. Numerical solution the the Smith–Hutton case for $\rho/\Gamma = 10^2$.

Figures 5.6 and 5.7 show the numerical solution to the Smith–Hutton case for $\rho/\Gamma = 10^4$ and 10^9 respectively. There is no apparent difference between two solution, what induces to think that for $\rho/\Gamma > 10^4$ the solution stays approximately the same. There are apparent discrepancies between the cases $\rho/\Gamma = 10^2$ and $\rho/\Gamma = 10^4$. In the latter transport clearly takes over diffusion, as the several color bands have approximately the same width, meaning diffusion has much less strength than transport.

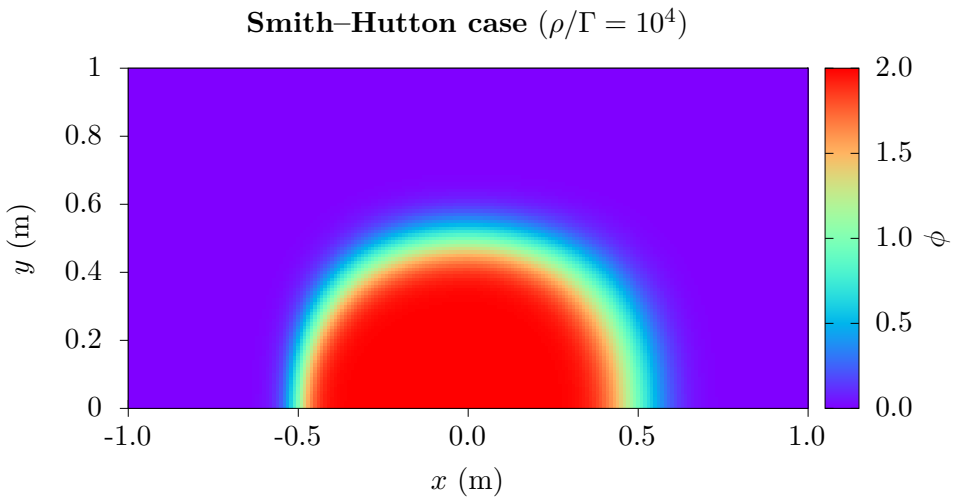


Figure 5.6. Numerical solution the the Smith–Hutton case for $\rho/\Gamma = 10^4$.

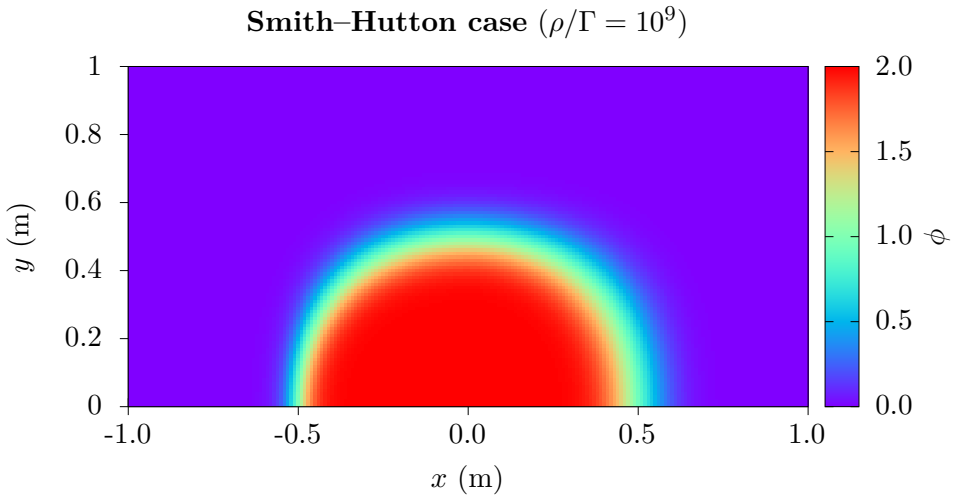


Figure 5.7. Numerical solution the the Smith–Hutton case for $\rho/\Gamma = 10^9$.

Figures 5.8 and 5.9 show the solution to the Smith–Hutton case for $\rho/\Gamma = 10^{-1}$ and $\rho/\Gamma = 10^{-9}$, respectively. Although a quotient $\rho/\Gamma = 10^{-9}$ implies transport is much weaker than for $\rho/\Gamma = 10^{-1}$, the differences between both solution are difficult to detect. The only apparent discrepancy is the central zone of Ω, where the transitions for $\rho/\Gamma = 10^{-9}$ are much smoother than for $\rho/\Gamma = 10^{-1}$, which implies transport has lost strength.

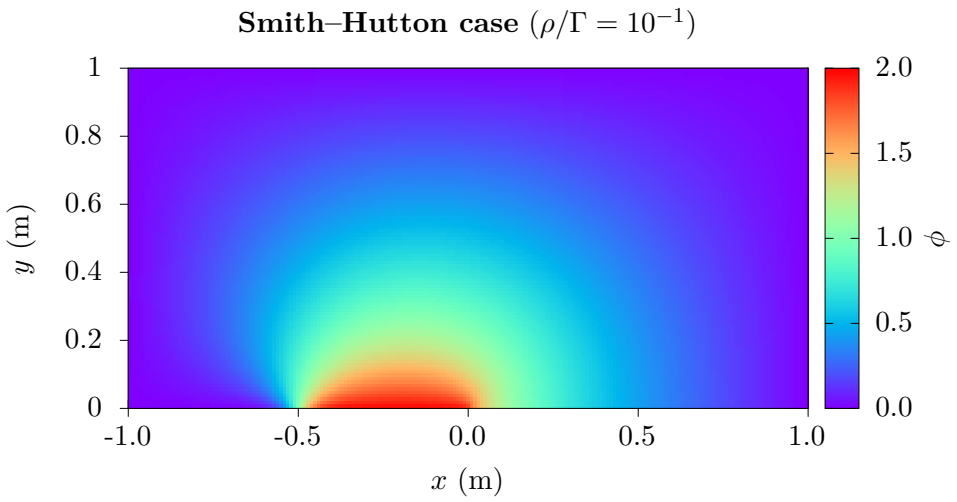


Figure 5.8. Numerical solution the the Smith–Hutton case for $\rho/\Gamma = 10^{-1}$.

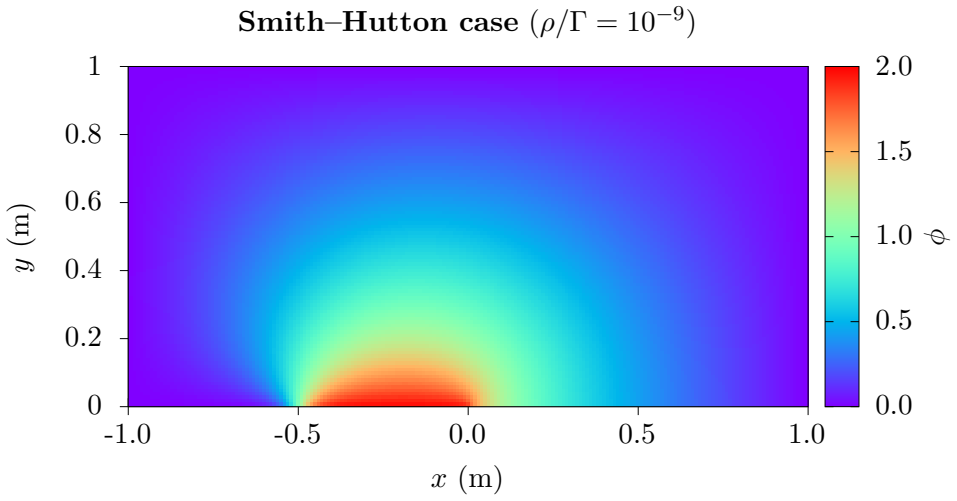


Figure 5.9. Numerical solution the the Smith–Hutton case for $\rho/\Gamma = 10^{-9}$.

References

[1] Sandro Salsa. *Partial Differential Equations in Action*. 1st ed. Springer, 2009.

[2] Wilfred Kaplan. *Advanced Calculus*. 5th ed. Pearson, 2002. Chap. 4, pp. 253–256.

[3] Pijush K Kundu, Ira M Cohen, and D Dowling. *Fluid Mechanics*. 6th ed. Elsevier, 2016. Chap. 3, pp. 99–102.

[4] CTTC. "Numerical resolution of the generic convection diffusion equation". Notes of the Course on Numerical Methods in Heat Transfer and Fluid Dynamics. July 2021.

[5] Lawrence C. Evans. *Partial Differential Equations*. 1st ed. Vol. 19. American Mathematical Society, 1998.

[6] Suhas V Patankar. *Numerical heat transfer and fluid flow*. 1st ed. McGraw–Hill Book Company, 1980. Chap. 5, pp. 79–111.

[7] Joel H Ferziger, Milovan Perić, and Robert L Street. *Computational methods for fluid dynamics*. 4th ed. Springer, 2002. Chap. 2, pp. 23–40.

[8] Joel H Ferziger, Milovan Perić, and Robert L Street. *Computational methods for fluid dynamics*. 4th ed. Springer, 2002. Chap. 4, pp. 81–110.

[9] Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical mathematics*. 1st ed. Vol. 37. Springer, 2000. Chap. 8, pp. 327–370.

[10] PH Gaskell and AKC Lau. "Curvature-compensated convective transport: SMART, a new boundedness-preserving transport algorithm". In: *International Journal for numerical methods in fluids* 8.6 (1988), pp. 617–641.

[11] Joel H Ferziger, Milovan Perić, and Robert L Street. *Computational methods for fluid dynamics*. 4th ed. Springer, 2002. Chap. 5, pp. 111–156.

[12] RM Smith and AG Hutton. "The numerical treatment of advection: A performance comparison of current methods". In: *Numerical Heat Transfer, Part A Applications* 5.4 (1982), pp. 439–461.

[13] Walter Rudin. *Real and Complex Analysis*. 3rd ed. McGraw–Hill, 1987.

[14] Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*. PUV, 2008. Chap. 3, pp. 59–99.

[15] José M. Mazón Ruiz. *Cálculo Diferencial. Teoría y problemas.* Vol. 140. American Mathematical Society, 2012. Chap. 2, pp. 33–58.

[16] Pau Martín de la Torre. "Equacions diferencials ordinàries". Notes de l'assignatura d'Equacions diferencials ordinàries impartida a la FME–UPC. 2020.

[17] H. Golub Gene and Charles F. Van Loan. *Matrix computations*. 4th ed. The Johns Hopkins University Press, 2013.

A Measure Theory

In this appendix we gather

A.1 Measurable spaces and measurable functions

In this appendix, X will denote an arbitrary set.

Definition. A σ–algebra of sets of X is a family $\mathcal{A}$ of subsets of X, $\mathcal{A} \subset \mathcal{P}(X)$, such that:

(i) $X \in \mathcal{A}$,

(ii) $A \in \mathcal{A}$ if and only if $A^c = X \setminus A \in \mathcal{A}$,

(iii) If $\{A_n\}_{n\geq 1} \subset \mathcal{A}$ is a countable family of sets, then $\cup_{n\geq 1} A_n \in \mathcal{A}$.

We say the pair $(X, \mathcal{A})$ is a measurable space.

As examples, both $\{\varnothing, X\}$ and $\mathcal{P}(X)$ are σ–algebras on X. The former is the trivial σ–algebra and is of no great interest. The latter is the power set of X, which is usually too big to work in. In many cases, it is convenient for a σ–algebra to contain certain subsets of X. This is made precise in the following definition:

Definition. The σ–algebra generated by a set $Z \subset \mathcal{P}(X)$ is

$$\langle Z \rangle = \bigcap_{\substack{\mathcal{A}\ \sigma\text{–algebra} \\ Z \subset \mathcal{A}}} \mathcal{A}$$

There exists many σ–algebras on $\mathbb{R}$, nonetheless we are interested on the smallest one containing all the open intervals. This is known as Borel's σ–algebra.

Definition. Borel's σ–algebra, denoted by $\mathcal{B}$, is the σ–algebra generated by $Z = \{(a, b) \subset \mathbb{R} \mid a < b\}$. The elements of $\mathcal{B}$ are called Borel's sets.

Definition. Let $(X, \mathcal{A})$ be a measurable space. A function $f\colon X \to \mathbb{R}$ is said to be $\mathcal{A}$–measurable (or measurable if $\mathcal{A}$ is clear) if $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$. We denote by $\mathcal{M}(X, \mathcal{A})$ the set of measurable functions.

In order to define Lebesgue's integral, it is convenient to consider functions which take infinite value at certain points, as well as the positive and negative parts of a function. To do so, we need the following definitions:

Definition.

(i) The extended real line is $\mathbb{R}^* = \mathbb{R} \cup \{-\infty, +\infty\}$.

(ii) The extended Borel σ–algebra is defined to be

$$\mathcal{B}^* = \{B,\, B \cup \{+\infty\},\, B \cup \{-\infty\},\, B \cup \{-\infty, +\infty\}\}_{B \in \mathcal{B}}$$

The definition of measurable function extends naturally to $\mathbb{R}^*$, that is to say, if $(X, \mathcal{A})$ is a measurable space, a function $f\colon X \to \mathbb{R}^*$ is said to be measurable whenever $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}^*$.

Definition. Let $(X, \mathcal{A})$ be a measurable space. The indicator function of a set $E \in \mathcal{A}$ is

$$\mathbb{I}_E(x) = \begin{cases} 1 & \text{if } x \in E \\ 0 & \text{if } x \notin E \end{cases}$$

Definition. Let $(X, \mathcal{A})$ be a measurable space and $f\colon X \to \mathbb{R}$. We define the positive and negative parts of f point by point as

$$f^+(x) = \max\{f(x), 0\}$$
$$f^-(x) = \max\{-f(x), 0\}$$

Proposition A.1. Let $(X, \mathcal{A})$ be a measurable space.

(1) The indicator function of a set $E \in \mathcal{A}$ is measurable.

(2) If $f\colon X \to \mathbb{R}$ is measurable, then f^+ and f^- are measurable.

In the real line $\mathbb{R}$ our natural notion of measure is the length. For instance, for $a, b \in \mathbb{R}$, we say the interval (a, b), has length $b - a$. This is also true for $(a, b]$, $[a, b)$ and $[a, b]$. In $\mathbb{R}^2$ and in $\mathbb{R}^3$ our notions of measure are surface and volume, respectively. For $\mathbb{R}^n$, although not so obvious, the notion of measure is the n–dimensional volume. The following definition generalizes the notion of measure to any space suitable enough to be measured.

Definition. A measure on a measurable space $(X, \mathcal{A})$ is a mapping $\mu\colon \mathcal{A} \to \mathbb{R}^*$ such that:

(i) $\mu(\varnothing) = 0$.

(ii) $\mu(A) \geq 0$ for all $A \in \mathcal{A}$.

(iii) σ–additivity: if $\{A_i\}_{i\geq 1} \subset \mathcal{A}$ is a countable family of measurable sets and $A_i \cap A_j = \varnothing$ whenever $i \neq j$, then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i\geq 1} \mu(A_i)$$

$(X, \mathcal{A}, \mu)$ is said to be a measure space.

The natural notion of n–dimensional volume in $\mathbb{R}^n$ that we have is given by Lebesgue's measure λ_n. Take $a_1 \leq b_1, \ldots, a_n \leq b_n$ and consider the intervals $[a_1, b_1], \ldots, [a_n, b_n] \subset \mathbb{R}$. Then the n–dimensional rectangle $\prod_{i=1}^n [a_i, b_i] = [a_1, b_1] \times \cdots \times [a_n, b_n]$ is a subset of $\mathbb{R}^n$ and its n–dimensional volume is

$$\lambda_n\left(\prod_{i=1}^n [a_i, b_i]\right) = \prod_{i=1}^n (b_i - a_i)$$

Assume $f\colon U \to \mathbb{R}$ is a continuous function defined on a compact subset U of $\mathbb{R}^n$, so that its integral over U is finite and given by

$$\int_U f \,\mathrm{d}x = \int_U f(x_1, \ldots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n$$

The notation $\mathrm{d}x_1 \cdots \mathrm{d}x_n$, besides telling us with respect to what variables the integral is being computed, it says how we are measuring in U. This way of measuring is the one given by Lebesgue's measure. We shall not discuss the construction of λ_n on $\mathbb{R}^n$. The only thing we need in order to continue is its uniqueness, that is to say, Lebesgue's measure on $\mathbb{R}^n$ is the only measure that gives the usual notion of n–dimensional volume.

Definition. Let $(X, \mathcal{A}, \mu)$ be a measure space.

(i) A set $E \in \mathcal{A}$ is said to be a null set if $\mu(E) = 0$.

(ii) Let P be a proposition. We say P is true μ–almost everywhere (μ–a.e.) if there exists $N \in \mathcal{A}$ such that $\mu(N) = 0$ and P is true on $N^c = X \setminus N$.

A.2 Lebesgue's integral

Let us begin by remembering the following theorem:

Theorem A.2 (Lebesgue's theorem)**.** Let $A = \prod_{i=1}^{n}[a_i, b_i] \subset \mathbb{R}^n$ be a compact rectangle and $f\colon A \to \mathbb{R}$ a bounded function. Then f is Riemann–integrable in A if, and only if, $\operatorname{disc}(f) = \{x \in U \mid f \text{ is not continuous at } x\}$ has zero λ_n measure.

Lebesgue's theorem asserts that a function is integrable in the sense of Riemann if the set where it is discontinuous is not too big (in the sense made precise previously). This is not true for Lebesgue's integral, that is to say, a function can be non–where continuous and integrable in Lebesgue's sense. Although not so computationally useful as Riemann's integral, it is a extremely powerful tool in advanced analysis.

Let $(X, \mathcal{A}, \mu)$ be a measure space and consider the set of positive measurable functions $\mathcal{M}^+(X, \mathcal{A}) = \{f \in \mathcal{M}(X, \mathcal{A}) \mid f \geq 0 \text{ for all } x \in X\}$. The idea behind Lebesgue's integral for functions in $\mathcal{M}^+(X, \mathcal{A})$ is to approximate them by functions $\varphi\colon X \to \mathbb{R}$ taking only finitely many values.

Definition. A function $\varphi\colon X \to \mathbb{R}$, $\varphi \in \mathcal{M}^+(X, \mathcal{A})$, is said to be simple if $|\operatorname{Im}\varphi| < \infty$.

Observe that every simple function can be written in the form $\varphi = \sum_{i=1}^{n} a_i \mathbb{I}_{E_i}$, where $n = |\operatorname{Im}\varphi|$, $a_1, \ldots, a_n \geq 0$ and $E_1, \ldots, E_n \in \mathcal{A}$. Nonetheless, it is convenient to have a canonical representation for simple functions.

Definition. The canonical representation of a simple function $\varphi \in \mathcal{M}^+(X, \mathcal{A})$ is

$$\varphi = \sum_{j=1}^{r} b_i \mathbb{I}_{A_i} \quad \text{such that} \quad \begin{cases} A_i \cap A_j = \varnothing & \text{if } i \neq j \\ b_i \neq b_j & \text{if } i \neq j \\ b_i \neq 0 & \text{for } i = 1, \ldots, r \end{cases}$$

Using the previous definition, we can define Lebesgue's integral for simple functions:

Definition. Let $\varphi\colon X \to \mathbb{R}$, $\varphi \in \mathcal{M}^+(X, \mathcal{A})$, be a simple function with canonical representation

$$\varphi = \sum_{j=1}^{r} b_i \mathbb{I}_{A_i}$$

Then we define Lebesgue's integral for φ by

$$\int_X \varphi \, \mathrm{d}\mu = \sum_{j=1}^{r} b_j \mu(A_j)$$

where we set $b_j \mu(A_j) = 0$ if $b_j = 0$ for any $\mu(A_j) \in [0, \infty]$.

Definition. Let $f \in \mathcal{M}^+(X, \mathcal{A})$. We define the Lebesgue integral of f over X to be

$$\int_X f \, \mathrm{d}\mu = \sup_{\substack{\varphi \leq f \\ \varphi \text{ simple}}} \int_X \varphi \, \mathrm{d}\mu$$

If $E \in \mathcal{A}$, then $f\mathbb{I}_E \in \mathcal{M}^+(X, \mathcal{A})$ and we define the Lebesgue integral of f over E by

$$\int_E f \, \mathrm{d}\mu = \int_X f\mathbb{I}_E \, \mathrm{d}\mu$$

Once Lebesgue's integral is defined for functions $\mathcal{M}^+(X, \mathcal{A})$, we would like to extend it to functions taking both positive and negative values.

Definition. We say a function $f \colon X \to \mathbb{R}$ is integrable if:

(i) f is a measurable function,

(ii) Both f^+ and f^-, which are measurable functions, have finite integral.

If f is an integrable function, we define the Lebesgue integral of f over X by

$$\int_X f \, \mathrm{d}\mu = \int_X f^+ \, \mathrm{d}\mu - \int_X f^- \, \mathrm{d}\mu$$

and

$$\int_E f \, \mathrm{d}\mu = \int_X f\mathbb{I}_E \, \mathrm{d}\mu = \int_E f^+ \, \mathrm{d}\mu - \int_E f^- \, \mathrm{d}\mu$$

whenever $E \in \mathcal{A}$. We denote the set of integrable functions by $L(X, \mathcal{A}, \mu)$.

A.3 L^p spaces

Let $(X, \mathcal{A}, \mu)$ be a measure space. Our aim is to define a norm on $L(X, \mathcal{A}, \mu)$ that makes it into a complete space, i.e. a space where every Cauchy sequence has limit. Recall that a norm on a $\mathbb{R}$–vector space F is a mapping $\|\cdot\| \colon F \to \mathbb{R}$ that satisfies the following:

(i) Non–negative: $\|x\| \geq 0$ for all $x \in F$.

(ii) Non–degenerate: $\|x\| = 0$ if and only if $x = 0$.

(iii) Product by scalars: $\|\lambda x\| = |\lambda| \|x\|$ for all $x \in F$ and $\lambda \in \mathbb{R}$.

(iv) Triangular inequality: $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in F$.

We begin by defining the mapping

$$\begin{aligned} N_\mu \colon L(X, \mathcal{A}, \mu) &\longrightarrow \mathbb{R} \\ f &\longmapsto \int_X |f| \, \mathrm{d}\mu \end{aligned}$$

which satisfies the following:

(1) $N_\mu(f) \geq 0$.

(2) $N_\mu(f) < \infty$ because f is integrable.

(3) $N_\mu(\alpha f) = |\alpha| N_\mu(f)$ for all $\alpha \in \mathbb{R}$.

(4) $N_\mu(f+g) \leq N_\mu(f) + N_\mu(g)$.

Nonetheless, $N_\mu(f) = 0$ does not imply $f \equiv 0$ in X. The reason for this is that f may take values different from 0 in null sets, that is to say, $N_\mu(f) = 0$ as long as $f = 0$ μ–a.e. We say that N_μ is a seminorm. Observe that redefining a function $f \in L(X, \mathcal{A}, \mu)$ in a null set does not change $\int_X f \, \mathrm{d}\mu$. Therefore, in order to make N_μ into a norm, we shall relate the functions in $L(X, \mathcal{A}, \mu)$ that differ only in null sets.

Definition. In $L(X, \mathcal{A}, \mu)$ we define the relation

$$f \sim_\mu g \iff f = g \ \mu\text{–a.e.} \tag{A.1}$$

Proposition A.3. Relation (A.1) is an equivalence relation in $L(X, \mathcal{A}, \mu)$.

Proof. Reflexivity and symmetry are trivially satisfied by $\sim_\mu$. Regarding the transitivity, assume $f \sim_\mu g$ and $g \sim_\mu h$. Let $N_1 \in \mathcal{A}$ be the null set where f and g differ. Similarly, let $N_2 \in \mathcal{A}$ be the null set where g and h differ. Then f and h differ in a measurable set $N \subset N_1 \cup N_2$, thus $\mu(N) \leq \mu(N_1 \cup N_2) \leq \mu(N_1) + \mu(N_2) = 0$, which implies $f \sim_\mu h$. □

Finally we can start defining the celebrated L^p spaces:

Definition. We define the space $L^1(X, \mathcal{A}, \mu)$ or Lebesgue space by

$$L^1(X, \mathcal{A}, \mu) = L(X, \mathcal{A}, \mu) \Big/ \sim_\mu$$

When both $\mathcal{A}$ and μ are clear, we shall write $L^1(X, \mathcal{A}, \mu) = L^1(X)$.

Definition. On $L^1(X, \mathcal{A}, \mu)$ we define the L^1–norm as

$$\|f\|_{L^1(X)} = \|f\|_1 = \int_X |f| \, \mathrm{d}\mu$$

Proposition A.4. If $[f] \in L^1(X)$ and $g \in [f]$, then $\int_X |f| \, \mathrm{d}\mu = \int_X |g| \, \mathrm{d}\mu$.

Proof. Let $N \in \mathcal{A}$ be the null set where f and g differ, i.e. $f \neq g$ in $N \in \mathcal{A}$ and $\mu(N) = A$. Let $M = X \setminus N$, then

$$\begin{aligned}\int_X |f| \, \mathrm{d}\mu &= \int_X |f|(\mathbb{I}_M + \mathbb{I}_N) \, \mathrm{d}\mu = \int_X |f|\mathbb{I}_M \, \mathrm{d}\mu + \int_X |f|\mathbb{I}_N \, \mathrm{d}\mu = \int_X |f|\mathbb{I}_M \, \mathrm{d}\mu \\ &= \int_X |g|\mathbb{I}_M \, \mathrm{d}\mu = \int_X |g|\mathbb{I}_M \, \mathrm{d}\mu + \int_X |g|\mathbb{I}_N \, \mathrm{d}\mu = \int_X |g| \, \mathrm{d}\mu\end{aligned}$$

□

Theorem A.5. $(L^1(X), \|\cdot\|_1)$ is a normed vector space.

Proof. $\|\cdot\|_1$ satisfies the non–negative, the product by scalars and the triangular inequality properties since N_μ does, thus it only remains to prove the non–degenerate property. Let $[f] \in L^1(X)$ such that $\|f\|_1 = 0$, then $|f| = 0$ μ–a.e., which is equivalent to $f = 0$ μ–a.e., which impies $f \in [0]$, so $\|\cdot\|_1$ is non–degenerate. □

The elements of $L^1(X)$ are the equivalence classes of $L(X, \mathcal{A}, \mu)$ given by the equivalence relation (A.1). However, proposition A.4 tells us that the integral does not depend on the representative of the class, hence we shall become less formal and treat the elements of $L^1(X)$ as functions.

Now we shall study the generalization of Lebesgue's space. Let $p \in (1, \infty)$ and consider the measurable functions $f\colon X \to \mathbb{R}$ such that

$$\int_X |f|^p \, \mathrm{d}\mu < \infty$$

We will denote by $L^p(X, \mathcal{A}, \mu)$ the set of equivalence classes of functions $|f|^p \in L(X, \mathcal{A}, \mu)$ with the equivalence relation $\sim_\mu$. When both $\mathcal{A}$ and μ are clear from the context, we shall write $L^p(X)$ in place of $L^p(X, \mathcal{A}, \mu)$.

Definition. On $L^p(X)$ we define the L^p–norm by

$$\|f\|_{L^p(X)} = \|f\|_p = \left(\int_X |f|^p \, \mathrm{d}\mu \right)^{1/p}$$

for $f \in L^p(X)$.

The difficult task on demonstrating that $\|\cdot\|_p$ is proving the triangular inequality. In the context of L^p spaces, this is known as Minkowski's inequality:

Theorem A.6 (Minkowski's inequality)**.** Let $p \geq 1$ and $f, g \in L^p(X)$. Then $f + g \in L^p(X)$ and $\|f + g\|_p \leq \|f\|_p + \|g\|_p$.

Theorem A.7 (Riesz–Fischer)**.** $(L^p(X), \|\cdot\|_p)$ is a complete space for $1 \leq p < \infty$.

Theorem A.6 tells us that $L^p(X)$, endowed with the L^p–norm, is a normed vector space, whereas A.7 asserts that is is complete. Hence $(L^p(X), \|\cdot\|_p)$ is a complete normed vector space, i.e. a Banach space, so we can do analysis on it.

In the definition of $L^p(X)$, we have asked for $1 \leq p < \infty$. Nonetheless, when $p = \infty$, we encounter the $L^\infty(X)$ space, formed by the essentially bounded functions, i.e. functions that are bounded almost everywhere except maybe on a null set. The L^∞–norm on $L^\infty(X)$ is given by

$$\|f\|_{L^\infty(X)} = \|f\|_\infty = \inf \{C \geq 0 \mid |f| \leq C \ \mu\text{–almost everywhere}\}$$

As it could be anticipated, $(L^\infty(X), \|\cdot\|_\infty)$ is also a Banach space.

A.4 Differentiation under the integral sign

In this appendix we gather two important theorems needed to justify some steps in the derivation of conservation laws in section 1. Despite these results are basic, a previous study of real analysis is required in order to understand and prove them. A good reference for the interested reader is Real and Complex Analysis of Walter Rudin [13].

A.5 Differentiation under the integral sign

Differentiation under the integral sign allows us to compute the derivative of an integral of a function of two parameters in a simple way. It is needed, for instance, when the mass conservation law or the heat diffusion equation are derived.

Let $(X, \mathcal{A}, \mu)$ be a measure space and let $[a, b] \subset \mathbb{R}$. Hereinafter we deal with functions $f\colon X \times [a, b] \to \mathbb{R}$, where $t \in [a, b]$ is the parameter on which f depends. We assume that $f(\cdot, t)$ is a measurable function for each $t \in [a, b]$.

Theorem A.8 (Differentiation under the integral sign)**.** Let $F(t) = \int_X f(x, t)\,\mathrm{d}\mu$. Assume that

(i) $f(x, t_0)$ is an integrable function for some $t_0 \in [a, b]$.

(ii) $\dfrac{\partial f}{\partial t}(x, t)$ is defined for all $(x, t) \in X \times [a, b]$.

(iii) There exists an integral function $g\colon X \to \mathbb{R}$ such that $\left|\dfrac{\partial f}{\partial t}(x, t)\right| \leq g(x)$ for all $(x, t) \in X \times [a, b]$.

Then F is a differentiable function and

$$F'(t) = \frac{\mathrm{d}}{\mathrm{d}t}F(t) = \int_X \frac{\partial f}{\partial t}(x, t)\,\mathrm{d}\mu$$

For the applications needed in this project, $X = \mathbb{R}^m$ with $1 \leq m \leq 3$, $\mathcal{A}$ is the Borel σ–algebra on $\mathbb{R}^m$ and μ is Lebesgue's measure on $\mathbb{R}^m$, which for most of the "natural" sets of $\mathcal{A}$ coincides with the usual notion of m–dimensional volume.

A.6 Lebesgue's differentiation lemma

A common way to derive a conservation law is to integrate some functions in a control volume, then apply Differentiation under the integral sign to obtain an integral equation and finally get to a differential equation using Lebesgue's differentiation lemma.

An intuitive way to understand and to motivate Lebesgue's differentiation lemma is the following. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a continuous function, let $a \in \mathbb{R}$ be a fixed point and let $F(x) = \int_a^x f(y)\,\mathrm{d}y$, which is a differentiable function. Due to a corollary of the Fundamental Theorem of Calculus, we have $F'(x) = f(x)$. Using the definition of derivative,

$$F'(x) = \lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = \lim_{h\to 0} \frac{1}{h}\left\{\int_a^{x+h} f(y)\,\mathrm{d}y - \int_a^x f(y)\,\mathrm{d}y\right\} = \lim_{h\to 0} \frac{1}{h}\int_x^{x+h} f(y)\,\mathrm{d}y = f(x)$$

Notice that the integral is divided by the length of the interval $[x, x + h]$, otherwise the limit would be zero. Lebesgue's lemma generalizes the previous equality by considering functions $f\colon \mathbb{R}^n \to \mathbb{R}$ and integrating them on open balls $B(x_0, r) = \{x \in \mathbb{R}^n \mid \|x - x_0\| < r\}$. Furthermore, the integral is divided by the n–dimensional volume of $B(x_0, r)$, which is denoted by $|B(x_0, r)|$.

Theorem A.9 (Lebesgue's differentiation lemma [5])**.** Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a locally integrable function.

(1) Then for almost everywhere point $x_0 \in \mathbb{R}^n$,

$$\frac{1}{|B(x_0,r)|}\int_{B(x_0,r)} f(x)\,\mathrm{d}x \to f(x_0) \quad \text{as } r \to 0$$

(2) In fact, for almost everywhere point $x_0 \in \mathbb{R}^n$,

$$\frac{1}{|B(x_0,r)|}\int_{B(x_0,r)} |f(x)-f(x_0)|\,\mathrm{d}x \to 0 \quad \text{as } r \to 0$$

B Sobolev spaces

C Ordinary Differential Equations

In this section we present a central theorem in basic Ordinary differential equations (ODE) theory regarding the existence and uniqueness of solution to initial value problems involving ODEs.

Recall that an ordinary differential equation is an equation

$$g(t, x(t), x'(t), \ldots, x^{(n)}(t)) = 0 \tag{C.1}$$

where the unknown $x(t) = (x_1(t), \ldots, x_m(t))^\top$ is a function of m components and a variable $t \in \mathbb{R}$, $x' = \dfrac{\mathrm{d}x}{\mathrm{d}t}$ and $g(t, y_1, \ldots, y_{n+1})$ with $y_1, \ldots, y_{n+1} \in \mathbb{R}^m$ is a function of $1 + m(n+1)$ variables.

C.1 General theory

We begin with the definition of a Lipschitz function:

Definition C.1. A function $f\colon \Omega \subset \mathbb{R}^n \to \mathbb{R}^m$ is said to be Lipschitz or Lipschitz continuous if there exists a constant L such that

$$\|f(x) - f(y)\| \leq L\|x - y\| \tag{C.2}$$

for every $x, y \in \Omega$. The constant L is called the Lipschitz constant of f [1].

Every Lipschitz function on Ω is also a continuous function in the usual sense on Ω. The converse is not true in general, and depends on Ω. The Lipschitz continuity is actually a very restrictive condition, since it imposes that the function can grow at most as a linear function.

Definition C.2. Let $(E, \|\cdot\|_E)$ and $(F, \|\cdot\|_F)$ be two finite dimensional normed vector spaces and let $A\colon E \to F$ be a linear mapping between them. The norm of A is defined to be

$$\|A\| := \sup\{\|Ax\|_F \mid x \in E,\ \|x\|_E \leq 1\} \tag{C.3}$$

It is a well known fact that $\|A\|$ is finite. In some cases, proving that a function is Lipschitz continuous is a laborious task. In these situations we have the following theorems:

Theorem C.3 (Corollary of the mean value theorem [14])**.** Let $\Omega \subset \mathbb{R}^n$ an open set and let $f\colon \Omega \subset \mathbb{R}^n \to \mathbb{R}^m$ a differentiable function on Ω. Let $x, y \in \Omega$ such that the segment $\overline{xy} = \{\lambda x + (1-\lambda)y \mid \lambda \in [0,1]\}$ is contained in Ω. Then the following inequality holds:

$$\|f(x) - f(y)\| \leq \sup_{z \in \overline{xy}} \|Df(z)\|\|x - y\| \tag{C.4}$$

Proof. See [14], page 78. □

Theorem C.4. Let $\Omega \subset \mathbb{R}^n$ a compact convex set and let $f \equiv (f_1, \ldots, f_m)\colon \Omega \subset \mathbb{R}^n \to \mathbb{R}^m$ a $\mathcal{C}^1(\Omega)$ function. Then f is Lipschitz in Ω.

Proof. The differential of Df, given by

$$Df = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} \tag{C.5}$$

is a linear mapping $Df\colon \mathbb{R}^n \to \mathbb{R}^m$. Since f is $\mathcal{C}^1(\Omega)$ function, the partial derivatives $\frac{\partial f_1}{\partial x_1}, \ldots, \frac{\partial f_m}{\partial x_n}$ are continuous functions on Ω. Moreover, the norm of the differential depends continuously on the partial derivatives $\frac{\partial f_1}{\partial x_1}, \ldots, \frac{\partial f_m}{\partial x_n}$ and on z. As a consequence, $\|Df\|$ is a continuous function on Ω. By Weierstrass theorem, $\|Df\|$ reaches a maximum M on Ω. Applying theorem C.3 we have

$$\|f(x) - f(y)\| \leq \sup_{z \in \overline{xy}} \|Df(z)\| \|x - y\| \leq M\|x - y\| \tag{C.6}$$

for all $x, y \in \Omega$, hence f is Lipschitz on $\overline{\Omega}$. □

Finally we can state the Picard–Lindelöf theorem. Let U be an open subset of $\mathbb{R}^{n+1}$ and let $f \in C(U, \mathbb{R})$, $(t_0, x_0) \in U$. Consider the following IVP:

$$\begin{cases} \dot{x}(t) = f(t, x) \\ x(t_0) = x_0 \end{cases} \tag{C.7}$$

The Picard–Lindelöf theorem gives us the existence and uniqueness of solution for (C.7).

Theorem C.5 (Picard–Lindelöf [15])**.** Suppose $f \in \mathcal{C}(U, \mathbb{R}^n)$, where U is an open subset of $\mathbb{R}^{n+1}$, and $(t_0, x_0) \in U$. If f is locally Lipschitz continuous in the second argument, uniformly with respect to the first, then there exists a unique local solution $x(t) \in \mathcal{C}^1(I)$ of the initial value problem (C.7), where I is some interval around t_0.

More specifically, if $V = [t_0, t_0 + T] \times \overline{B(x_0, \delta)} \subset U$ and M denotes the maximum of $|f|$ on V, then the solution exists at least for $t \in [t_0, t_0 + T_0]$ and remains in $\overline{B(x_0, \delta)}$ where $T_0 = \min\left\{T, \frac{\delta}{M}\right\}$. The analogous result holds for the interval $[t_0 - T, t_0]$.

Proof. See [15], page 38. □

C.2 Linear equations

Let $\mathcal{M}_{n\times n}(\mathbb{R})$ and $\mathcal{M}_{n\times n}(\mathbb{C})$ denote the space of $n \times n$ matrices with real or complex coefficients, respectively. Recall that an ODE or a system of ODEs is linear if it has the form

$$\dot{x} = A(t)x + b(t) \tag{C.8}$$

where $A(t) \in \mathcal{M}_{n\times n}(\mathbb{R})$ (or $A(t) \in \mathcal{M}_{n\times n}(\mathbb{C})$) and $b(t) \in \mathbb{R}^n$ for all $t \in I \subset \mathbb{R}$. The following theorem establishes the existence and uniqueness of initial value problems for linear ODEs.

Theorem C.6. Let $I \subset \mathbb{R}$ be an open interval and let $A \in \mathcal{C}(I, \mathcal{M}_{n\times n}(\mathbb{R}))$ and $b \in \mathcal{C}(I, \mathbb{R}^n)$. Then for all $(t_0, x_0) \in I \times \mathbb{R}^n$, the initial value problem

$$\begin{cases} \dot{x} = A(t)x + b(t) \quad t \in I \\ \dot{x}(t_0) = x_0 \end{cases} \tag{C.9}$$

has a unique solution $\varphi\colon I \to \mathbb{R}^n$.

Proof. [16] □

In order for theorem C.6 to be useful for the project, we shall show that any n–th order linear ODE

$$x^{(n)} + a_{n-1}(t)\, x^{(n-1)} + \cdots + a_1(t)\, x' + a_0(t)\, x = b(t) \tag{C.10}$$

can be casted into a linear system of ODEs such as (C.8). We define the functions

$$y_1 = x,\ y_2 = x', \ldots,\ y_{n-1} = x^{(n-2)},\ y_n = x^{(n-1)} \tag{C.11}$$

thus

$$\begin{cases} y_1' = x' = y_2 \\ y_2' = x'' = y_3 \\ \quad \vdots \\ y_{n-1}' = x^{(n-1)} = y_n \\ \begin{aligned} y_n' = x^{(n)} &= -a_0(t)\, x - a_1(t)\, x' - \cdots - a_{n-1}(t)\, x^{(n-1)} + b(t) \\ &= -a_0(t)\, y_1 - a_1(t)\, y_2 - \cdots - a_{n-1}(t)\, y_n + b(t) \end{aligned} \end{cases} \tag{C.12}$$

which in matrix form is rewritten as:

$$\begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_{n-1}' \\ y_n' \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0(t) & -a_1(t) & -a_2(t) & \cdots & a_{n-1}(t) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ b(t) \end{pmatrix} \tag{C.13}$$

As long as $a_0, a_1, \ldots, a_{n-1}, b\colon I \subset \mathbb{R} \to \mathbb{R}$ are continuous functions, theorem C.6 can be applied to system (C.13).

D Numerical resolution of linear systems

In this section we present several iterative methods and the LU decomposition, all of them used to solve linear systems. The reference is [17].

Let $\mathcal{M}_{n\times n}(\mathbb{R})$ be the space of $n \times n$ matrices with real coefficients. We consider linear systems of the form $Ax = b$, where $A \in \mathcal{M}_{n\times n}(\mathbb{R})$ is the system matrix with $\det(A) \neq 0$ and $b \in \mathbb{R}^n$ is the vector of independent terms. Recall the following definitions:

Definition D.1. Let $A = (a_{ij}) \in \mathcal{M}_{n\times n}(\mathbb{R})$ be a matrix. Then:

(i) A is symmetric if $a_{ij} = a_{ji}$ for all $1 \leq i, j \leq n$.

(ii) A is positive definite if $c^\top Ac > 0$ for all $c \in \mathbb{R}^n \setminus \{0\}$.

(iii) A is diagonally dominant if

$$|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ij}|, \quad 1 \leq i \leq n \tag{D.1}$$

and strictly diagonally dominant if the previous inequality is strict.

D.1 Iterative methods

Consider the linear system $Ax = b$ with $A = (a_{ij}) \in \mathcal{M}_{n\times n}(\mathbb{R})$ a non–singular matrix, $b \in \mathbb{R}^n$ the vector of independent terms and $x = A^{-1}b \in \mathbb{R}^n$ the solution. Iterative methods for linear systems generate a sequence $\{x^{(k)}\}_{k\geq 0} \subset \mathbb{R}^n$ that ideally converge to the solution, i.e. $\lim_{k\to\infty} x^{(k)} = x$.

D.1.1 Jacobi's method

Let $x^{(k)}$ be the current approximation of $x = A^{-1}b$ and assume $a_{ii} \neq 0$ for all $1 \leq n \leq n$. The first idea for an iterative method is to compute $x^{(k+1)}$ as

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{n} a_{ij} x_j^{(k)} \right), \quad 1 \leq i \leq n \tag{D.2}$$

The method given by (D.2) is known as Jacobi's method. We have the following theorem about its convergence.

Theorem D.2. If $A \in \mathcal{M}_{n\times n}(\mathbb{R})$ is strictly diagonally dominant, the Jacobi's method converges to $x = A^{-1}b$.

Proof. See [17], Ch. 11, pg. 615. □

D.1.2 Gauss–Seidel's method

Notice that before prior to calculate $x_i^{(k+1)}$ in (D.2), the components $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$ have had to be computed. Since these are already available once $x_i^{(k+1)}$ is being calculated, a natural improvement to Jacobi's method is the following:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right), \quad 1 \leq i \leq n \tag{D.3}$$

Recall that A is a positive definite matrix if $c^\top Ac > 0$ for all $c \in \mathbb{R}^n \setminus \{0\}$. We say that A is a diagonally dominant matrix if it satisfies

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad 1 \leq i \neq n \tag{D.4}$$

The following theorem gives us two sufficient conditions so that Gauss–Seidel's method converges.

Theorem D.3. If A is a positive definite symmetric matrix or it is strictly diagonally dominant by rows, then the Gauss–Seidel converges to the linear system solution.

Proof. See [17], Ch. 11, pg. 615. □

D.1.3 Relaxation method

From Gauss–Seidel's method we have

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) = x_i^{(k)} + \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(k)} \right) \tag{D.5}$$

which means that $x_i^{(k+1)}$ is equal to $x_i^{(k)}$ plus a correction. The correction can be multiplied by a constant ω

$$x_i^{(k+1)} = x_i^{(k)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(k)} \right), \quad 1 \leq i \leq n \tag{D.6}$$

so as to accelerate the convergence. The constant ω is known as relaxation constant and the method given by (D.6) is known as relaxation method. In general, there are no results providing an optimal ω to improve the convergence velocity.

D.1.4 Stop criterion

An iterative method must stop at some point once the iteration $x^{(k)}$ is close enough to the solution x. The distance between $x^{(k)}$ and x is given by $\left\| x^{(k)} - x \right\|$. There are many norms in $\mathbb{R}^n$, thus a natural question is which one to use. The following theorem asserts that the election of norm is not relevant:

Theorem D.4. Any two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on $\mathbb{R}^n$ are equivalent, i.e. there exist constants $c, C > 0$ such that

$$c\|x\|_a \leq \|x\|_b \leq C\|x\|_b \tag{D.7}$$

for all $x \in \mathbb{R}^n$.

Hence any two norms on $\mathbb{R}^n$ provide the same notion of distance. A common choice for iterative methods due to its low computational cost is the supremum norm, given by

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i| \tag{D.8}$$

To halt the iteration, one normally controls the norm $\left\| x^{(k+1)} - x^{(k)} \right\|$. Given a constant $\delta > 0$ small enough so that the approximation of x is good, the iteration is stopped when $\left\| x^{(k+1)} - x^{(k)} \right\| < \delta$.

D.2 LU decomposition

Let U be an upper triangular non–singular matrix, that is to say, a matrix of the form

$$U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1,n-1} & u_{1n} \\ 0 & u_{22} & \cdots & u_{2,n-1} & u_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & u_{n-1,n-1} & u_{n-1,n} \\ 0 & 0 & \cdots & 0 & u_{nn} \end{pmatrix}, \; u_{ii} \neq 0 \quad 1 \leq i \leq n \tag{D.9}$$

or equivalently

$$U = (u_{ij})_{i,j=1 \div n} = \begin{cases} u_{ij} = 0 & \text{if } i > j \\ u_{ij} \neq 0 & \text{if } i = j \\ u_{ij} \in \mathbb{R} & \text{otherwise} \end{cases} \tag{D.10}$$

Let $b \in \mathbb{R}^n$. The the linear system $Ux = b$ can be easily solved with the backward substitution algorithm

$$x_{n-i} = \frac{1}{u_{n-i,n-i}} \left(b_{n-i} - \sum_{j=n-i+1}^{n} u_{n-i,j}\, x_j \right), \quad 0 \leq i \leq n-1 \tag{D.11}$$

Let L be a lower triangular non–singular matrix,

$$L = \begin{pmatrix} \ell_{11} & 0 & \cdots & 0 & 0 \\ \ell_{21} & \ell_{22} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \ell_{n-1,1} & \ell_{n-1,2} & \cdots & \ell_{n-1,n-1} & 0 \\ \ell_{n1} & \ell_{n,2} & \cdots & \ell_{n,n-1} & \ell_{nn} \end{pmatrix}, \ \ell_{ii} \neq 0 \quad 1 \leq i \leq n \tag{D.12}$$

or which is the same,

$$L = (\ell_{ij}) = \begin{cases} \ell_{ij} = 0 & \text{if } i < j \\ \ell_{ij} \neq 0 & \text{if } i = j \\ \ell_{ij} \in \mathbb{R} & \text{otherwise} \end{cases} \tag{D.13}$$

The system $Lx = b$ can be solved with the forward substitution algorithm

$$x_i = \frac{1}{\ell_{ii}} \left(b_i - \sum_{j=1}^{i-1} \ell_{ij}\, x_j \right), \quad 1 \leq i \leq n \tag{D.14}$$

Theorem D.5. Let $A \in \mathcal{M}_{n\times n}(\mathbb{R})$.

(1) Assume that there exist a lower triangular matrix L and an upper triangular matrix U such that $A = LU$. Then L and U are unique.

(2) If the k–th principal minor of A is non–null for all $1 \leq k \leq n$, that is to say, if

$$\begin{vmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{vmatrix} \neq 0 \tag{D.15}$$

then there exist L and U such that $A = LU$.

(3) If the Gaussian elimination can be carried out on A, then $A = LU$ where

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ \ell_{21} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \ell_{n-1,1} & \ell_{n-1,2} & \cdots & 1 & 0 \\ \ell_{n1} & \ell_{n,2} & \cdots & \ell_{n,n-1} & 1 \end{pmatrix} \qquad U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1,n-1} & u_{1n} \\ 0 & u_{22} & \cdots & u_{2,n-1} & u_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & u_{n-1,n-1} & u_{n-1,n} \\ 0 & 0 & \cdots & 0 & u_{nn} \end{pmatrix}$$

Certain matrices are not suitable for Gaussian elimination. In such cases it is necessary to permute the columns of A during the elimination. This may be expressed with an invertible permutation matrix P, so that A is decomposed as $PA = LU$ or $A = P^{-1}LU$. Once A is decomposed, the linear system $PAx = b$ can be solved as two triangular systems with the previously seen algorithms. Indeed, since $PAx = LUx = b$, the systems to be solved are $Ly = b$ and $Ux = y$.