ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Christophe Ancey

Laboratoire hydraulique environnementale (LHE)

École Polytechnique Fédérale de Lausanne
Écublens

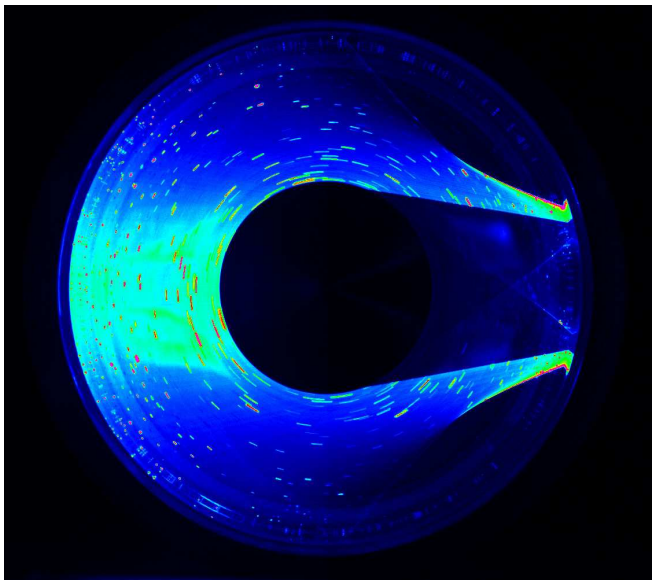
CH-1015 Lausanne

Lecture Notes

Fluid Dynamics and Rheology

---

draft 1.9 of 29th November 2009



C. Ancey,
EPFL, ENAC/IIC/LHE,
Ecublens, CH-1015 Lausanne, Suisse
christophe.ancey@epfl.ch, lhe.epfl.ch

Fluid Dynamics and Rheology / C. Ancey

La gestion typographique a été réalisée à l'aide du package *french.sty* de Bernard Gaulle.

Remerciements : à Sébastien Wiederseiner, Martin Rentschler, et Nicolas Andreini pour la relecture du manuscrit.

« Le physicien ne peut demander à l'analyste de lui révéler une vérité nouvelle ; tout au plus celui-ci pourrait-il l'aider à la pressentir. Il y a longtemps que personne ne songe plus à devancer l'expérience, ou à construire le monde de toutes pièces sur quelques hypothèses hâtives. De toutes ces constructions où l'on se complaisait encore naïvement il y a un siècle, il ne reste aujourd'hui plus que des ruines. Toutes les lois sont donc tirées de l'expérience, mais pour les énoncer, il faut une langue spéciale ; le langage ordinaire est trop pauvre, elle est d'ailleurs trop vague, pour exprimer des rapports si délicats, si riches et si précis. Voilà donc une première raison pour laquelle le physicien ne peut se passer des mathématiques ; elles lui fournissent la seule langue qu'il puisse parler. »

Henri Poincaré, in *La valeur de la science*

The physician cannot ask the analyst to reveal a new truth; at best the analyst could help him to have a feel of it. It has been a long time that nobody has not anticipated the experience or built the world from scratch on a few hasty assumptions. All these constructions where people still naively delighted a century ago, are today in ruins. All laws are drawn from experience, but expressing them requires a special language; usual language is too poor, it is also too vague to express relations that are so delicate, so rich and precise. That is the main reason why physicists cannot live without mathematics; it provides them the only language they can speak.

Henri Poincaré, in *The Value of Science*

Table des matières

Notations, formulas, & Conventions

The following notations and rules are used:

- Vectors, matrices, and tensors are in bold characters.
- For mathematical variables, I use slanted fonts.
- Functions, operators, and dimensionless numbers are typed using a Roman font.
- The symbol O (capital O) indicates that there is a one-sided bound, e.g., $u = O(v)$ means that the limit of u/v exists and is finite (neither zero nor infinity). On many occasions, it means "is of the order of", but not always (so be careful).
- The symbol o is a shorthand notation to for 'much smaller than', e.g. $u = o(v)$ means that $u \ll v$.
- I do not use the notation D/Dt to refer to the material derivative (also called convective or Lagrangian derivative), but $\mathrm{d}/\mathrm{d}t$ (that must not be confused with ordinary time derivative). I believe that the context is mostly sufficient to determine the meaning of the differential operator.
- The symbol $\propto$ means "proportional to".
- The symbol $\sim$ or $\approx$ means "nearly equal to".
- I use units of the international system (IS): meter [m] for length, second [s] for time, and kilogram [kg] for mass. Units are specified by using square brackets.
- For the computations with complex numbers, I use $\Re$ to refer to the real part of a complex and $\imath$ is the imaginary number.
- The superscript † after a vector/tensor means the transpose of this vector/tensor.
- We use $\mathbf{1}$ to refer to the unit tensor (identity tensor/matrix).
- Einstein's convention means that when summing variables, we omit the symbol $\sum$ and we repeat the indice. For instance we have $\mathbf{a} \cdot \mathbf{b} = a_i b_i$.
- The gradient operator is denoted by the *nabla* symbol ∇. The divergence of any scalar or tensorial quantity f is denoted by $\nabla \cdot f$. For the Laplacian operator, I indifferently use ∇^2 or $\triangle$. The curl of any vector $\mathbf{v}$ is denoted by $\nabla \times \mathbf{v}$. We can use the following rule to check the consistency of an operator

Operation name	*Operator symbol*	*Order of result*
gradient	∇	$\Sigma + 1$
divergence or outer product	$\nabla\cdot$	$\Sigma - 1$
curl	$\nabla\times$	Σ
Laplacian	∇^2	Σ

- The scalar product of two vectors $\mathbf{a}$ and $\mathbf{b}$ is denoted by $\mathbf{a} \cdot \mathbf{b}$. The dyadic or tensor product of $\mathbf{a}$ and $\mathbf{b}$ is denoted by $\mathbf{ab}$. The product between a tensor $\mathbf{A}$ and a vector $\mathbf{a}$ is denoted by $\mathbf{A} \cdot \mathbf{a}$. The cross product of two vectors $\mathbf{a}$ and $\mathbf{b}$ is denoted by $\mathbf{a} \times \mathbf{b}$. The inner product of two tensors is denoted by the double dot ":" (keep in mind that for second-order tensors $\mathbf{a}$ and $\mathbf{b}$ we have $\mathbf{a} : \mathbf{b} = \mathrm{tr}\,\mathbf{ab}$). We can use the following rule to check the consistency of a multiplication

Operation name	*Multiplication sign*	*Order of result*
dyadic or tensorial product	none	Σ
cross or outer product	$\times$	$\Sigma - 1$
scalar or inner product	$\cdot$	$\Sigma - 2$
double product	:	$\Sigma - 4$

  Recall that the order of a scalar is 0, a vector is of order 1, and a tensor is of order at least 2. For instance, if $\mathbf{a}$ and $\mathbf{b}$ denotes vectors and $\mathbf{T}$ is a tensor, $\mathbf{T} \cdot \mathbf{a}$ is order $2 + 1 - 2 = 1$.

- The gradient of a vector $\mathbf{a}$ is a tensor $\nabla\mathbf{a}$, whose components in a Cartesian frame x_i are
$$\frac{\partial a_j}{\partial x_i}.$$
  The divergence of a second-order tensor M_{ij} is a vector $\nabla \cdot \mathbf{M}$, whose jth component in a Cartesian frame is
$$\frac{\partial M_{ij}}{\partial x_i}.$$

- The tensorial product of two vectors $\mathbf{a}$ and $\mathbf{b}$ provides a tensor $\mathbf{ab}$ such that for any vector $\mathbf{c}$, we have $(\mathbf{ab})\mathbf{c} = (\mathbf{b} \cdot \mathbf{c})\mathbf{a}$.

- A vector field such that $\nabla \cdot \mathbf{v} = \mathbf{0}$ is said to be *solenoidal*. A function f satisfying the Laplace equation $\nabla^2 f = 0$ is said to be *harmonic*. A function f such that $\nabla^4 f = 0$ is said to be *biharmonic*. A vectorial field $\mathbf{v}$ such $\nabla \times \mathbf{v} = 0$ is said to be *irrotational*.

- An extensive use is made of the *Green-Ostrogradski* theorem (also called the *divergence* theorem):
$$\int_{\mathcal{V}} \nabla \cdot \ \mathbf{u}\, \mathrm{d}\mathcal{V} = \int_{\mathcal{S}} \mathbf{u} \cdot \mathbf{n}\, \mathrm{d}\mathcal{S},$$
  where $\mathcal{S}$ is the surface bounding the volume $\mathcal{V}$ and $\mathbf{n}$ is the unit normal to the infinitesimal surface $\mathrm{d}\mathcal{S}$. A closely related theorem for scalar quantities f is
$$\int_{\mathcal{V}} \nabla f\, \mathrm{d}\mathcal{V} = \int_{\mathcal{S}} f\mathbf{n}\, \mathrm{d}\mathcal{S}.$$

- For some algebraic computations, we need to use
  - Cartesian coordinates (x, y, z); or
  - spherical coordinates $(x = r\cos\varphi\sin\theta)$, $y = r\sin\theta\sin\varphi$, $z = r\cos\theta)$ with $0 \leq \theta \leq \pi$ and $-\pi \leq \varphi \leq \pi$, $\mathrm{d}S = r^2\sin\theta\mathrm{d}\theta\mathrm{d}\varphi$ on a sphere of radius r, $\mathrm{d}V = r^2\sin\theta\mathrm{d}r\mathrm{d}\theta\mathrm{d}\varphi$.

- Some useful formulas on vector and tensor products

$$\mathbf{N} : \mathbf{M} = \mathbf{M} : \mathbf{N},$$

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c},$$

$$(\mathbf{M} \cdot \mathbf{a}) \cdot \mathbf{b} = \mathbf{M} : (\mathbf{ab}) \text{ and } \mathbf{a} \cdot (\mathbf{b} \cdot \mathbf{M}) = \mathbf{M} : (\mathbf{ab}),$$

$$\mathbf{ab} : \mathbf{cd} = \mathbf{a} \cdot (\mathbf{b} \cdot \mathbf{cd}) = \mathbf{a} \cdot ((\mathbf{b} \cdot \mathbf{c})\mathbf{d}) = (\mathbf{a} \cdot \mathbf{b})(\mathbf{c} \cdot \mathbf{d}) = \mathbf{ac} : \mathbf{bd}$$

$$\nabla(fg) = g\nabla f + f\nabla g,$$

$$\nabla \cdot (f\mathbf{a}) = \mathbf{a} \cdot \nabla f + f\nabla \cdot \mathbf{a},$$

$$\nabla \cdot (\mathbf{a} \times \mathbf{b}) = \mathbf{b}(\nabla \times \mathbf{a}) - \mathbf{a}(\nabla \times \mathbf{b}),$$

$$\nabla \cdot \nabla\mathbf{a} = \frac{1}{2}\nabla(\mathbf{a} \cdot \mathbf{a}) - \mathbf{a} \times (\nabla \times \mathbf{a}),$$

$$\nabla \cdot \mathbf{ab} = \mathbf{a}\nabla\mathbf{b} + \mathbf{b}\nabla \cdot \mathbf{a}$$

$$\mathbf{1} : \nabla\mathbf{a} = \nabla \cdot \mathbf{a},$$

$$\nabla \cdot (f\mathbf{1}) = \nabla f,$$

- and on derivatives

$$(\mathbf{a} \cdot \nabla)\mathbf{b} = \mathbf{a} \cdot (\nabla\mathbf{b})^T,$$

$$\frac{\partial f(x)}{\partial \mathbf{x}} = \frac{\mathbf{x}}{x}\frac{\partial f(x)}{\partial x},$$

$$\mathbf{ab} : (\nabla\mathbf{c}) = \mathbf{a} \cdot (\mathbf{b}\nabla)\,\mathbf{c},$$

  with $x = |\mathbf{x}|$.

- For some computations, we need the use the *Dirac function* $\int_{\mathbb{R}^3} \delta(\mathbf{x})\mathrm{d}\mathbf{x} = 1$, $\int_{\mathbb{R}^3} \delta(\mathbf{x} - \mathbf{x}_0)g(\mathbf{x})\mathrm{d}\mathbf{x} = g(\mathbf{x}_0)$, and $\delta(\mathbf{x}) = -\nabla^2(4\pi x)^{-1} = -\nabla^4 x(8\pi)^{-1}$, where $x = |\mathbf{x}|$. The last two expressions are derived by applying the Green formula to the function $1/x$ (see any textbook on distributions).

- The *Fourier transform* in an n-dimensional space is defined as

$$\hat{f}(\boldsymbol{\xi}) = \int_{\mathbb{R}^n} f(\mathbf{x})e^{-\imath\boldsymbol{\xi}\cdot\mathbf{x}}\mathrm{d}\mathbf{x},$$

  for any continuous function. Conversely, the inverse Fourier transform is defined as

$$f(\mathbf{x}) = \int_{\mathbb{R}^n} \hat{f}(\mathbf{x})e^{\imath\boldsymbol{\xi}\cdot\mathbf{x}}\mathrm{d}\boldsymbol{\xi}.$$

Further reading

These lecture notes gives an overview of tools and concepts in fluid dynamics and rheology through a series of different problems of particular relevance to free-surface flows. For each topic considered here, we will outline the key elements and point the student toward the most helpful references and authoritative works. The student is also referred to available books introducing rheology (Barnes, 1997; Tanner, 1988) for a more complete presentation; the tutorials written by Middleton & Wilcock (1994) on mechanical and rheological applications in geophysics and by Barnes (2000) provide a shorter introduction to rheology.

Continuum Mechanics, rheology

- P. Chadwick, *Continuum Mechanics, Concise Theory and Problems*, (Dover, New York) 187 p.
- K. Hutter and K. Jöhnk, *Continuum Methods of Physical Modeling* (Springer, Berlin, 2004) 635 p.
- H.A. Barnes, J.F. Hutton and K. Walters, *An introduction to rheology* (Elsevier, Amsterdam, 1997).
- H.A. Barnes, *A Handbook of Elementary Rheology* (University of Wales, Aberystwyth, 2000).
- K. Walters, *Rheometry* (Chapman and Hall, London, 1975).
- D.V. Boger and K. Walters, *Rheological Phenomena in Focus* (Elsevier, Amsterdam, 1993) 156 p.
- B.D. Coleman, H. Markowitz and W. Noll, *Viscometric flows of non-Newtonian fluids* (Springer-Verlag, Berlin, 1966) 130 p.
- C. Truesdell, *Rational Thermodynamics* (Springer Verlag, New York, 1984).
- C. Truesdell, The meaning of viscometry in fluid dynamics, *Annual Review of Fluid Mechanics*, **6** (1974) 111–147.

Mathematical skills

- Zwillinger, D., Handbook of differential equations, 775 pp., Academic Press, Boston, 1992.
- King, A.C., J. Billingham, and S.R. Otto, *Differential Equations: Linear, Nonlinear, Ordinary, Partial*, 541 pp., Cambridge University Press, Cambridge, 2003.

- Zauderer, E., *Partial Differential Equations of Applied Mathematics*, 779 pp., John Wiley & Sons, New York, 1983.
- Mei, C.C., *Mathematical Analysis in Engineering*, 461 pp., Cambridge University Press, Cambridge, 1995.
- Bender, C.M., and S.A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, 593 pp., Springer, New York, 1999.

Fluid mechanics

- S.B. Pope, *Turbulent Flows* (Cambridge University Press, Cambridge, 2000) 771 p.
- W. Zdunkowski and A. Bott, *Dynamics of the Atmosphere* (Cambridge University Press, Cambridge, 2003) 719 p.
- C. Pozrikidis, *Boundary Integral and Singularity Methods for Linearized Viscous Flows* (Cambridge University Press, Cambridge, 1992) 259 p.
- G.K. Batchelor, *An introduction to fluid dynamics* (Cambridge University Press, 1967) 614 p.
- H. Lamb, *Hydrodynamics* (Cambridge University Press, Cambridge, 1932).

1 Conservation laws

Within the framework of continuum mechanics, the details of the material microstructure are forgotten and we only examine the bulk behavior, e.g. how does a material deform when it experiences a given state of stress? Mathematically, we introduce the *constitutive equation*, which relates deformations and/or rates of deformation to stresses. Essentially continuum mechanics provides tools and rules that make it possible to:

- express constitutive equations in a proper form, i.e., in a tensorial form that satisfies the rules of physics;
- obtain equations that govern the bulk motion.

Most phenomenological laws we can infer from experiments are in a scalar form. For instance, in rheometry, the only information we obtain in most cases is the flow curve, i.e., the relation $\tau = f(\dot{\gamma})$, whereas we need more information to model the three-dimensional behavior of fluids. The question is how to express a three-dimensional constitutive equations. We shall see that the reply is not as easy as can be thought at first glance. We start with the Newtonian case, the simplest case that we can imagine. We then continue by reviewing the basic equations used in continuum mechanics (conservation of mass, momentum, and energy). Emphasis is then given to providing a few examples of application.

1.1 Why is continuum mechanics useful? An historical perspective

Newton's and Trouton's experiments were run on very viscous materials, but their interpretation, if cursorily made, leads to different values of viscosity.

1.1.1 Paradoxical experimental results?

In 1687, Isaac Newton proposed that "the resistance which arises from the lack of slipperiness of the parts of the liquid, other things being equal, is proportional to the velocity with which the parts of the liquid are separated from one another". This forms the basic statement behind the theory of Newtonian fluid mechanics. Translated into modern scientific terms, this sentence means that the resistance to flow (per unit area) τ is proportional to the velocity gradient U/h:

$$\tau = \mu \frac{U}{h}, \tag{1.1}$$

where U is the relative velocity with which the upper plate moves and h is the thickness of fluid separating the two plates (see Fig. 1.1). μ is a coefficient intrinsic to the material, which is termed *dynamic viscosity*. This relationship is of great practical importance for many reasons:

- it is the simplest way of expressing the constitutive equation for a fluid (linear behavior);
- it provides a convenient experimental method for measuring the constitutive parameter μ by measuring the shear stress exerted by the fluid on the upper plate moving with a velocity U (or conversely by measuring the velocity when a given tangential force is applied to the upper plate).

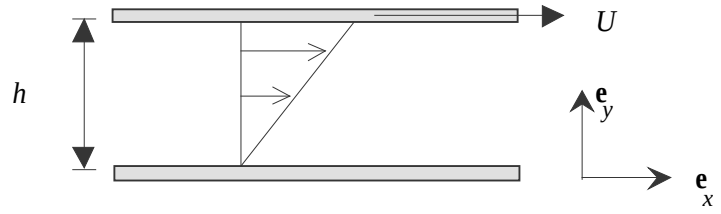


Figure 1.1. *Illustration of a fluid sheared by a moving upper plate*

In 1904, Trouton did experiments on mineral pitch involving stretching the fluid with a given velocity. Figure 1.2 depicts the principle of this experiment. The fluid undergoes a uniaxial elongation achieved with a constant elongation rate $\dot{\alpha}$, defined as the relative deformation rate: $\dot{\alpha} = \dot{l}/l$, where l is the fluid sample length. For his experiments, Trouton found a linear relationship between the applied force per unit area $\boldsymbol{\sigma}$ and the elongation rate:

$$\boldsymbol{\sigma} = \mu_e \alpha = \mu_e \frac{1}{l} \frac{\mathrm{d}l}{\mathrm{d}t}. \tag{1.2}$$

This relationship was structurally very similar to the one proposed by Newton but it introduced a new material parameter, which is now called *Trouton viscosity*. This constitutive

parameter was found to be three times greater than the Newtonian viscosity inferred from steady simple-shear experiments: $\mu_e = 3\mu$. At first glance, this result is both comforting since behavior is still linear (the resulting stress varies linearly with the applied strain rate) and disturbing since the value of the linearity coefficient depends on the type of experiment.

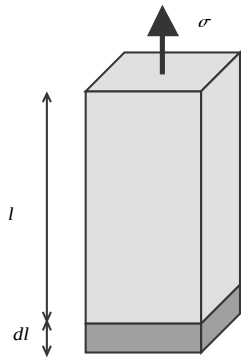


Figure 1.2. *Typical deformation of a material experiencing a normal stress σ*

1.1.2 How to remove the paradox?

In fact, Trouton's result does not lead to a paradox if we are careful to express the constitutive parameter in a tensorial form rather than a purely scalar form. This was achieved by Navier and Stokes, who independently developed a consistent three-dimensional theory for Newtonian viscous fluids. For a simple fluid, the stress tensor σ can be cast in the following form:

$$\sigma = -p\mathbf{1} + \mathbf{s}, \tag{1.3}$$

where p is called the fluid pressure and $\mathbf{s}$ is the extra-stress tensor representing the stresses resulting from a relative motion within the fluid. It is also called the *deviatoric stress tensor* since it represents the departure from equilibrium. The pressure term used in Eq. (1.3) is defined as (minus) the average of the three normal stresses $p = -\mathrm{tr}\,\boldsymbol{\sigma}/3$. This also implies that $(\mathrm{tr}\,\mathbf{s} = 0)$. The pressure used in Eq. (1.3) is analogous to the hydrostatic fluid pressure in the sense that it is a measure of the local intensity of the squeezing of the fluid. The connection between this purely mechanical definition and the term pressure used in thermodynamics is not simple. For a Newtonian viscous fluid, the Navier-Stokes equation postulates that the extra-stress tensor is linearly linked to the strain-rate tensor $\mathbf{d} = (\nabla\mathbf{u} + \nabla\mathbf{u}^\dagger)/2$:

$$\mathbf{s} = 2\eta\mathbf{d}, \tag{1.4}$$

where $\mathbf{u}$ is the local fluid velocity and η is called the *Newtonian viscosity*. It is worth noticing that the constitutive equation is expressed as a relationship between the extra-stress tensor and the local properties of the fluid, which are assumed to depend only on the instantaneous distribution of velocity (more precisely, on the departure from uniformity of that distribution). There are many arguments from continuum mechanics and analysis of molecular transport of momentum in fluids, which show that the local velocity gradient $\nabla\mathbf{u}$ is the parameter of the flow field with most relevance to the deviatoric stress. On the contrary, the pressure is not a constitutive parameter of the moving fluid. When the fluid is compressible, the pressure p can be inferred from the free energy, but it is indeterminate for incompressible Newtonian fluids.

If we return to the previous experiments, we infer from the momentum equation that the velocity field is linear : $\mathbf{u} = U\mathbf{e}_x y/h$. We easily infer that the shear rate is: $\dot{\gamma} = \partial u/\partial y = U/h$ and then comparing (1.4) to (1.1) leads to: $\eta = \mu$.

Thus, the Newtonian viscosity corresponds to the *dynamic viscosity*, measured in a simple-shear flow. In the case of a uniaxial elongation and when inertia can be neglected, the components of the strain-rate tensor are:

$$\mathbf{d} = \begin{bmatrix} \dot{\alpha} & 0 & 0 \\ 0 & -\dot{\alpha}/2 & 0 \\ 0 & 0 & -\dot{\alpha}/2 \end{bmatrix}. \tag{1.5}$$

At the same time, the stress tensor can be written as:

$$\boldsymbol{\sigma} = \begin{bmatrix} \sigma & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{1.6}$$

Comparing (1.3), (1.5), and (1.6) leads to: $p = -\eta\dot{\alpha}$ and $\sigma = 3\eta\dot{\alpha}$, that is: $\mu_e = 3\eta$, confirming that the Trouton elongational viscosity is three times greater than the Newtonian viscosity. It turns out that Trouton's and Newton's experiments reflect the same constitutive behavior. This example shows the importance of an appropriate tensorial form for expressing the stress tensor. In the present case, the tensorial form (1.4) may be seen as a simple generalization of the simple shear expression (1.1).

1.2 Fundamentals of Continuum Mechanics

1.2.1 Kinematics

It is customary to start a continuum mechanics course with the notion of Lagrangian and Eulerian descriptions:

- in an Eulerian description of the matter, attention is focused on what happens in a given volume control regardless of the history of particles contained in this volume;
- in a Lagrangian description, we follow up the motion of each particle that was at a given position at $t = 0$.

This duality in the matter description disappears very quickly from students' memory. Fluid mechanicians make use of Eulerian tools almost exclusively since for Newtonian fluids, the physics is governed by the past history, whereas solid mechanicians give preference to the Lagrangian description because for small deformations, there is no much differences between the descriptions.

However, in an advanced course on rheology and continuum mechanics, emphasis is given to the dual nature of materials which can exhibit both solid-like and fluid-like properties. Much attention has been brought to providing a unified vision of continuum mechanics that is sufficiently general to be applied to a wide range of rheological behaviors. This unified view extends the classic mechanics in several ways:

- large deformations to cope with viscoplasticity and viscoelasticity;
- thermodynamics to take irreversible processes into account.

Here we will focus our attention to a classic description of material deformation and the reader is referred to specialized books that expound more sophisticated theories of deformation (Bird *et al.*, 1987; Tanner, 1988; Morrison, 2001).

In the following, we shall following focus on Eulerian form of the equations of motion, but keep in mind that, especially in advanced fluid mechanics, Lagrangian representations of the equations are very useful (e.g., see Pope, 2000; Minier & Peirano, 2001; Zdunkowski & Bott, 2003, in the field of turbulence or atmospheric flows).

The gradient tensor

In order to characterize the deformation of a body, it is usually helpful to determine how neighboring points behave, i.e. how the increment $\mathrm{d}\mathbf{X}$ in an initial frame is transformed (see Fig. 1.3), including:

- stretching/contraction of length,
- rotation of elements due to solid rotation and shearing.

We take three neighboring points in the initial frame of reference $\mathcal{C}_0$, called A, B, and C, forming a right angle (see Fig. 1.3). Because of deformation, there may be

- stretching in a direction $\mathbf{e}_1$ or $\mathbf{e}_2$,

- solid rotation around a given axis with an angle α,
- pure shear, i.e. the angle $\theta_0 = \pi/2$ between $\mathbf{X}_1 = \mathbf{AB}$ and $\mathbf{X}_2 = \mathbf{BC}$ has been altered.

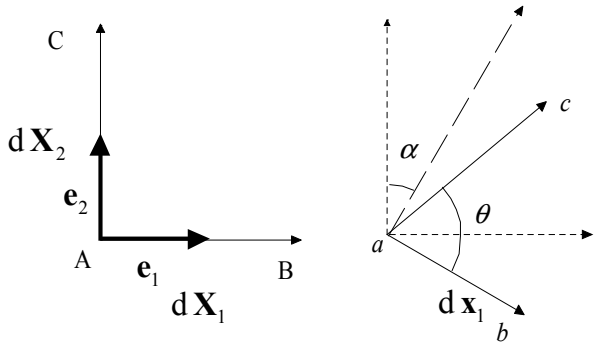


Figure 1.3. *Deformation of a right angle.*

For this purpose, we introduce the transformation

$$\mathrm{d}\mathbf{X} \to \mathrm{d}\mathbf{x},$$

where $\mathbf{x}(\mathbf{X},\ t)$ is the position occupied at time t by a particle that was earlier at $t = 0$ at the position $\mathbf{X}$ in the initial frame of reference $\mathcal{C}_0$. Positioning $\mathbf{x}$ can be achieved in the same frame of reference $\mathcal{C}_0$ or in another frame $\mathcal{C}(t)$. Differentiating the relation $\mathbf{x} = \mathbf{x}(\mathbf{X},\ t)$ leads to

$$\mathrm{d}\mathbf{x} = \mathbf{F} \cdot \mathrm{d}\mathbf{X},$$

where $\mathbf{F}$ is called the *gradient tensor*. When we use the same frame to refer to the current and initial configurations, we can introduce the displacement vector $\mathbf{u}$ such that

$$\mathbf{x} = \mathbf{u} + \mathbf{X},$$

from which we infer (after differentiation)

$$F_{ij} = \delta_{ij} + \frac{\partial u_i}{\partial X_j}.$$

The deformation of the angle (AB, AC) can be determined by using the scalar product

$$\mathrm{d}\mathbf{x}_1 \cdot \mathrm{d}\mathbf{x}_2 = (\mathbf{F} \cdot \mathrm{d}\mathbf{X}_1) \cdot (\mathbf{F} \cdot \mathrm{d}\mathbf{X}_2),$$

which can be transformed into

$$\mathrm{d}\mathbf{x}_1 \cdot \mathrm{d}\mathbf{x}_2 = \mathrm{d}\mathbf{X}_1 \cdot (\mathbf{C} \cdot \mathrm{d}\mathbf{X}_2),$$

where $\mathbf{C} = \mathbf{F}^\dagger \cdot \mathbf{F}$ is a symmetric tensor called the *stretch tensor* or *Cauchy-Green tensor*. The relative variation of the scalar product is then

$$\mathrm{d}\mathbf{x}_1 \cdot \mathrm{d}\mathbf{x}_2 - \mathrm{d}\mathbf{X}_1 \cdot \mathrm{d}\mathbf{X}_2 = 2\mathrm{d}\mathbf{X}_1 \cdot (\mathbf{E} \cdot \mathrm{d}\mathbf{X}_2),$$

where

$$\mathbf{E} = \frac{1}{2}(\mathbf{C} - \mathbf{1}) = \frac{1}{2}(\mathbf{F}^\dagger \cdot \mathbf{F} - \mathbf{1}),$$

is the *strain tensor of Green-Lagrange.*

Let us now express the stretching of a length increment $\mathrm{d}s = \sqrt{\mathrm{d}\mathbf{x}_1 \cdot \mathrm{d}\mathbf{x}_1}$

$$\mathrm{d}s_1 = (\mathbf{e}_1 \cdot \mathbf{C} \cdot \mathbf{e}_1)^{1/2} \mathrm{d}S,$$

where $\mathrm{d}S_1 = \sqrt{\mathrm{d}\mathbf{X}_1 \cdot \mathrm{d}\mathbf{X}_1}$. We deduce the relative stretching, i.e. the strain in direction $\mathbf{e}_1$

$$\epsilon_1 = \frac{\mathrm{d}s_1 - \mathrm{d}S_1}{\mathrm{d}S_1} = (\mathbf{e}_1 \mathbf{C} \mathbf{e}_1)^{1/2} - 1 = \sqrt{1 + 2E_{11}} - 1.$$

We can also characterize the angle θ

$$\cos\theta = \frac{\mathrm{d}\mathbf{x}_1 \cdot \mathrm{d}\mathbf{x}_2}{\mathrm{d}s_1 \mathrm{d}s_2} = \frac{2E_{12}}{\sqrt{1 + 2E_{11}}\sqrt{1 + 2E_{22}}},$$

which shows that

- the diagonal components of $\mathbf{E}$ give information on strains in the axis directions;
- the off-diagonal terms of $\mathbf{E}$ specify how a angle of an initially right wedge is deformed.

In order to get rid of the solid rotation that is not related to deformation, we make use of a theorem, called the *polar decomposition* theorem, that says that any tensor can be broken down in a unique way into an orthogonal[1] tensor $\mathbf{R}$ (representing block rotation) and symmetric (pure deformation) tensor [length-increment variation + angle variation]. Applied to the gradient tensor, this theorem allows us to write:

$$\mathbf{F} = \mathbf{R} \cdot \mathbf{U} = \mathbf{V} \cdot \mathbf{R},$$

where $\mathbf{U}$ (resp. $\mathbf{V}$) is the right (resp. left) pure strain tensor.

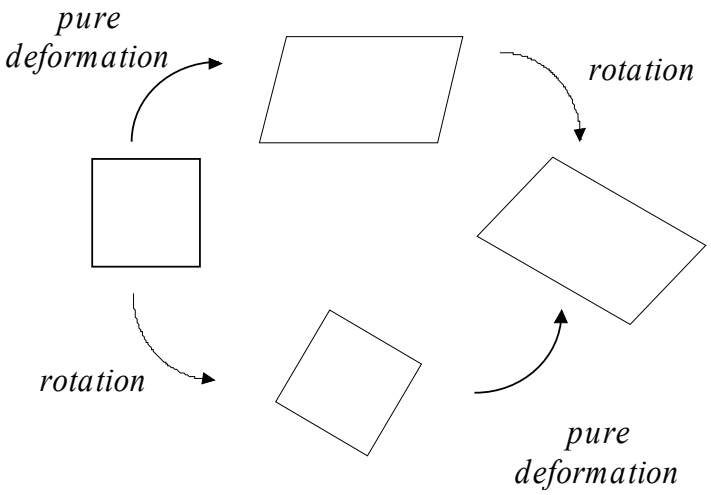


Figure 1.4. *Polar decomposition of the gradient tensor.*

This theorem can also be applied to other tensors such as $\mathbf{C}$ and $\mathbf{E}$, but in that case we can take benefit from the symmetry[2] of these tensors together with the symmetry in $\mathbf{U}$ (or

[1] Recall that if a tensor is orthogonal, then $\mathbf{R}^\dagger \cdot \mathbf{R} = \mathbf{R} \cdot \mathbf{R}^\dagger = \mathbf{1}$ and $\det \mathbf{R} = 1$.

[2] Recall that a real-valued symmetric tensor can be diagonalized and its eigenvector basis is orthogonal.

$\mathbf{V}$). In their eigenvector bases, deformation corresponds to length variations with no angle variation.

In addition to the length, we can also characterize the deformation of surfaces and volumes. The transformation of an infinitesimal volume $\mathrm{d}V$ in the initial frame is given by the Jacobian J of the transformation $\mathrm{d}\mathbf{X} \to \mathrm{d}\mathbf{x}$

$$\mathrm{d}v = J\mathrm{d}V \text{ with } J = |\det \mathbf{F}|.$$

Similarly, an oriented infinitesimal element of surface $\mathrm{d}\mathbf{A}$ can be expressed as $\mathrm{d}\mathbf{A} = \mathrm{d}\mathbf{X}_1 \times \mathrm{d}\mathbf{X}_2$, which is transformed into

$$\mathrm{d}\mathbf{a} = \mathrm{d}\mathbf{x}_1 \times \mathrm{d}\mathbf{x}_2 = J(\mathbf{F}^{-1})^\dagger \cdot \mathrm{d}\mathbf{A}.$$

♣ **Example**. – For instance, let us consider a shear strain in the form

$$x_1 = X_1 + \gamma(t)X_2,\ x_2 = X_2, \text{ and } x_3 = X_3,$$

where $\gamma(t)$ is called the shear amplitude and $(x_1,\ x_2,\ x_3)$ is the coordinates in a Cartesian frame. We find that this transformation keeps the volumes constant ($J = 1$) and

$$\mathbf{F} = \begin{bmatrix} 1 & \gamma & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},\ \mathbf{C} = \begin{bmatrix} 1 & \gamma & 0 \\ \gamma & 1+\gamma^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \text{ and } \mathbf{E} = \begin{bmatrix} 0 & \gamma/2 & 0 \\ \gamma/2 & \gamma^2/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

□

Small deformations

When deformations are small, we can linearize the tensors to get rid of nonlinear terms. Essentially this means that the displacement $\mathbf{u} = \mathbf{x} - \mathbf{X}$ between the current and initial positions is very small. We can deduce that

- there is no much differences between the frame $\mathcal{C}_0$ and $\mathcal{C}(t)$, which can then be assumed to be the same;
- we can write $\mathbf{F} = \mathbf{1} + \mathbf{H}$ with $\mathbf{H} = \nabla_X \mathbf{x}$ with "small" entries.

We deduce that the Cauchy-Green tensor can be linearized in the following way

$$\mathbf{C} = \mathbf{F}^\dagger \cdot \mathbf{F} = \mathbf{1} + \mathbf{H} + \mathbf{H}^\dagger = 1 + 2\boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ is called the (linearized) strain tensor and is defined as the symmetric part of $\mathbf{H}$

$$\boldsymbol{\epsilon} = \frac{1}{2}(\mathbf{H} + \mathbf{H}^\dagger) = \frac{1}{2}\left(\nabla_X \mathbf{x} + \nabla_X \mathbf{x}^\dagger\right) = \mathbf{E}.$$

Rate of strain

In order to characterize the velocity, we introduce the time derivative of $\mathbf{x}(\mathbf{X},\ t)$. In a Lagrangian system, $\mathbf{X}$ corresponds to the initial condition at $t = 0$, whereas in an Eulerian system, $\mathbf{X}$ is the position previously occupied by a particle and is a function of time.

The time derivative of an infinitesimal length increment is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{d}\mathbf{x} = \dot{\mathbf{F}} \cdot \mathrm{d}\mathbf{X} = \dot{\mathbf{F}} \cdot \mathbf{F}^{-1} \cdot \mathrm{d}\mathbf{x} = \mathbf{L} \cdot \mathrm{d}\mathbf{x},$$

where $\mathbf{L} = \dot{\mathbf{F}} \cdot \mathbf{F}^{-1} = \nabla_x \mathbf{v}$ is called the *velocity gradient tensor*. It is customary to break down into a symmetric and antisymmetric contribution

$$\mathbf{d} = \frac{1}{2}(\mathbf{L} + \mathbf{L}^\dagger) \text{ and } \mathbf{w} = \frac{1}{2}(\mathbf{L} - \mathbf{L}^\dagger).$$

The symmetric part $\mathbf{d}$ describes the strain rate and hence is called the *strain-rate tensor* (sometimes the stretching tensor), whereas the antisymmetric part $\mathbf{w}$ called the *vorticity tensor* (sometimes the spin tensor) corresponds to the rotational of the velocity field.

Recall that when a tensor $\mathbf{w}$ is antisymmetric, this implies the existence of a vector $\boldsymbol{\omega}$ such that for any vector $\mathbf{n}$, we have $\mathbf{w} \cdot \mathbf{n} = \boldsymbol{\omega} \times \mathbf{n}$. Let us expand the velocity $\mathbf{v}(\mathbf{x} + \mathrm{d}\mathbf{x})$

$$\mathbf{v}(\mathbf{x} + \mathrm{d}\mathbf{x}) = \mathbf{v}(\mathbf{x}) + \nabla_x \mathbf{v} \cdot \mathrm{d}\mathbf{x} + \cdots = \mathbf{v}(\mathbf{x}) + \mathbf{d}(\mathbf{x}) \cdot \mathrm{d}\mathbf{x} + \mathbf{w}(\mathbf{x}) \cdot \mathrm{d}\mathbf{x} + \cdots,$$

which means that to first order, we have

$$\mathbf{v}(\mathbf{x} + \mathrm{d}\mathbf{x}) = \mathbf{v}(\mathbf{x}) + \mathbf{d} \cdot \mathrm{d}\mathbf{x} + \boldsymbol{\omega} \times \mathrm{d}\mathbf{x}.$$

This means that the local variation in the velocity field can be broken down into a strain-rate contribution and another contribution corresponding to solid rotation

Refined kinematical description

Here we have deliberately ignored the issues related to the frames in which the tensor components are written. We have just introduced the configurations $\mathcal{C}_0$ and $\mathcal{C}(t)$ and any quantity can be defined in either configuration. Another approach is to introduce a curvilinear coordinate system that is linked to the material deformations: when the body deforms, the initial axes of $\mathcal{C}_0$ deform in the same way and define a system of material coordinates that is embedded in the material and deform with it, as shown in Figure 1.5.

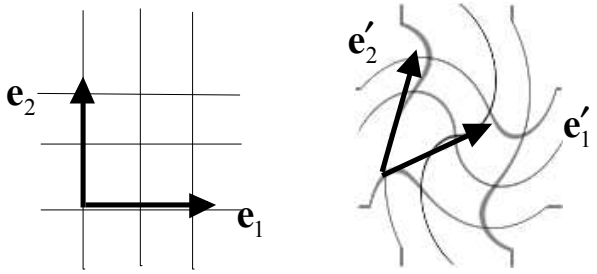


Figure 1.5. *Polar decomposition of the gradient tensor.*

The idea of a convected coordinate system embedded in a flowing film and deforming with it was developed by Oldroyd (1950) in the 1950s. This idea was motivated by the desire to formulate constitutive equations that are (i) independent of any frame of reference, (ii) independent of the position, and translational and rotational motion, of the element as a whole in space, and (iii) independent of the states of neighboring material elements.

1.2.2 Stress tensor

Definition of the stress tensor

We are going to introduce a new tensor called the *stress tensor*, which is used to compute the stresses exerted on an infinitesimal surface δS oriented by the outward unit vector $\mathbf{n}$. The stress $\boldsymbol{\tau}$ exerted on δS is defined as the ratio of the forces $\mathbf{f}$ per unit surface when δS becomes smaller and smaller:

$$\boldsymbol{\tau} = \lim_{\delta S \to 0} \frac{\mathbf{f}}{\delta S}.$$

Using Cauchy's lemma (force balance on a tetrahedron), it can be shown that there exists a tensor $\boldsymbol{\sigma}$, the *stress tensor*, such that

$$\boldsymbol{\tau} = \boldsymbol{\sigma} \cdot \mathbf{n}, \tag{1.7}$$

which means that the stress linearly varies with the normal $\mathbf{n}$. In a Cartesian frame, this tensor is represented by a symmetric matrix.

It is worth reminding that this construction is based on a postulate (force balance with no torque), which implies that other constructions are possible (e.g., micropolar or Cosserat medium) (Germain, 1973*a*,*b*). More generally, it is possible to use the virtual power principle to derive all the fundamental equations used in continuum mechanics and in that case, the stress tensor can formally be derived from the inner energy dissipation rate Φ

$$\sigma_{ij} = \frac{\partial \Phi}{\partial d_{ij}}. \tag{1.8}$$

This approach to continuum mechanics turns out to be fruitful for a number of problems in elasticity (membrum and shell theory) and in fluid mechanics. Indeed, on some occasions, it is easier to compute the energy dissipated in a system and, in that case, to compute the stress tensor using (1.8); for instance, a number of approximate computations of the bulk viscosity of a dilute particle suspension were done in this way (Einstein, 1911; Frankel & Acrivos, 1967).

Pressure and extra-stress

For a simple fluid at rest, the stress tensor reduces to

$$\boldsymbol{\sigma} = -p\mathbf{1},$$

where p is called the *pressure*. From the thermodynamic viewpoint, the pressure is a function of the density ρ and temperature T: $p = p(\rho,\ T)$. Making use of the (Helmholtz) free energy $F = E - TS$, with S entropy, E internal energy, then we can show

$$p = \rho^2 \frac{\partial F}{\partial \rho}.$$

When the density is constant (incompressible material) or the flow is isochoric, the thermodynamic definition of pressures is no longer valid and the *thermodynamic pressure* must be replaced by the *hydrodynamic pressure*. The latter is an undetermined function that can be specified on solving the governing equations with specific boundary conditions.

When an incompressible simple fluid at rest is slightly disturbed, we can imagine that the stress can be expressed as

$$\boldsymbol{\sigma} = -p\mathbf{1} + \mathbf{s},$$

where $\mathbf{s}$ is called the extra-stress tensor and represents the departure from the static equilibrium. We shall see that for a wide class of fluids, this extra-stress is a function of the strain-rate tensor alone $\mathbf{d}$, which leads to posing: $\boldsymbol{\sigma} = -p\mathbf{1} + \mathbf{s}(\mathbf{d})$. The simplest dependence of $\mathbf{s}$ on $\mathbf{d}$ that we can imagine is the linear relation: $\mathbf{s} = 2\mu\mathbf{d}$, i.e., the Newtonian constitutive equation. There is another motivation for writing the stress tensor as the sum of a pressure term and an additional contribution $\mathbf{s}$. Indeed, for an incompressible fluid, the mass balance imposes some constraints on the motion; there are internal stresses $\boldsymbol{\sigma}'$ that make the fluid incompressible. If these stresses are assumed to induce no energy dissipation, then for any deformation, we must have $\Phi = \mathrm{tr}\,(\boldsymbol{\sigma}\mathbf{d}) = 0$ or, in other words, $\boldsymbol{\sigma}' = -p\mathbf{1}$ since $\mathrm{tr}\,\mathbf{d} = 0$ (incompressible fluid). In the sequel, we shall see that there are some rules that must be satisfied in specifying a particular form of constitutive equation.

This way of expressing the stress tensor can be generalized to any type of material. It is customary in rheology to break down the stress tensor into two parts:

$$\boldsymbol{\sigma} = -p\mathbf{1} + \mathbf{s},$$

where p is now called the *mean pressure*

$$p = -\frac{1}{3}\mathrm{tr}\,\boldsymbol{\sigma},$$

and $\mathbf{s}$ is called the *deviatoric stress tensor* since it represents the departure from an equivalent equilibrium state.

Describing the rheological behavior of a material involves determining the relation between the stress tensor $\boldsymbol{\sigma}$ and the gradient tensor $\mathbf{F}$. The relation $\boldsymbol{\sigma} = \mathcal{F}(\mathbf{F})$, where $\mathcal{F}$ is a functional is called the *constitutive equation*. When the material is incompressible, the constitutive equation is usually defined with the extra-stress tensor since the stress tensor is defined to within an arbitrary function (pressure): $\mathbf{s} = \mathcal{F}(\mathbf{F})$.

Keep in mind that pressure may have a meaning that differs depending on the context: thermodynamic pressure, hydrodynamic pressure, mean pressure.

1.2.3 Admissibility of a constitutive equations

There are a number of rules that must be used to produce constitutive equations that are admissible from the rational and physical standpoints. The establishment of these rules have been the subject of long debates and has been approached in a number of ways, the interrelation of which are by no means easy to understand (Oldroyd, 1950; Truesdell, 1966, 1974). Here we will deal with general principles without expounding all the details.

Objectivity or material indifference

A constant behavior in physics is to express laws that do not depend on a particular system of reference. Let us try now to apply this fundamental principle to the formulation of constitutive equation. We assume that a constitutive equation is a mathematical relation between the stress and the deformation, using a short-hand notation, we express

$$\boldsymbol{\sigma} = \mathcal{F}(\mathbf{F})$$

where $\mathcal{F}$ is a functional depending on the gradient tensor $\mathbf{F}$. We have seen earlier that $\mathbf{F}$ can include solid rotation. Solid rotation as wells as translation of a reference or a change of clock

frame must not modify the physics, which implies that we must get rid of solid rotation and translation when expressing a constitutive equation.

Principle 1: *A constitutive equations is invariant under any change of reference frame.*

In practice this means that we make a change of variable

$$\mathbf{x}' = \mathbf{R} \cdot \mathbf{x} + \mathbf{b} \text{ and } t' = t + a,$$

which means that the image $\mathbf{x}'$ experiences a rotation (the tensor $\mathbf{R}$ can be time-dependent) and a translation ($\mathbf{b}$ is a constant vector) with respect to the original point $\mathbf{x}$; in addition there is change in the time reference (a being a constant). Then, for a quantity to be objective we must check that:

- a scalar field remains the same: $s'(\mathbf{x}',\ t') = s(\mathbf{x},\ t)$,
- a vector must satisfy $\mathbf{x}'(\mathbf{x}',\ t') = \mathbf{R} \cdot \mathbf{x}(\mathbf{x},\ t)$
- a tensor $\mathbf{T}$ is objective if it transforms objective vectors into objective vectors, i.e., $\mathbf{T}' \cdot \mathbf{x}' = \mathbf{R} \cdot \mathbf{T} \cdot \mathbf{x}$ or equivalently

$$\mathbf{T}'(\mathbf{x}',\ t') = \mathbf{R} \cdot \mathbf{T}(\mathbf{x},\ t) \cdot \mathbf{R}^{\dagger}.$$

The issue lies in the dependence of $\mathbf{F}$ on a particular frame. For instance, we introduce a rotation of the reference frame, the gradient tensor in the new frame is

$$F'_{ij} = \frac{\partial x'_i}{\partial X_j} = \frac{\partial x'_i}{\partial x_k}\frac{\partial x_k}{\partial X_j},$$

with $R_{ik} = \partial x'_i / \partial x_k$ an orthogonal tensor that corresponds to a frame rotation: $\mathbf{x}' = \mathbf{R} \cdot \mathbf{x}$ (see Figure 1.6). We deduce that $\mathbf{F}' = \mathbf{R} \cdot \mathbf{F}$, which is not admissible.

Now, if we make use of the Cauchy tensor

$$\mathbf{C}' = \mathbf{F}'^{\dagger} \cdot \mathbf{F}' = \mathbf{F}^{\dagger} \cdot \mathbf{R}^{\dagger} \cdot \mathbf{R} \cdot \mathbf{F} = \mathbf{F}^{\dagger} \cdot \mathbf{F} = \mathbf{C},$$

which shows that $\mathbf{C}$ is not an objective tensor since it does depend on the frame. Similarly, it can be shown that the strain tensor $\mathbf{E}$ and the strain-rate tensor $\mathbf{d}$ are objective.

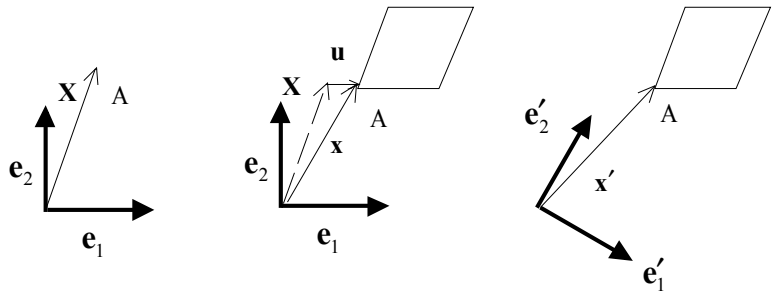


Figure 1.6. *Change of reference.*

Determinism

Principle 2: *The history of the thermo-kinetic process experienced by the material fully determines the current rheological and thermodynamic state of the material.* This principle must be relaxed slightly when the material is incompressible because the stress state is determined to within the hydrostatic pressure (which depends on the boundary conditions and the problem geometry).

Local action

Principle 3: *The thermodynamic process of a material at a given point is completely determined by the history of the thermo-kinetic process to which the neighborhood of the point was submitted.* In other words, the stress tensor at a given point does not depend on movements occurring at finite distance from this point.

1.2.4 Specific properties of material

A material is said to be

- *homogeneous* when the constitutive equation does not depend on the point considered;
- *isotropic* when the material response is invariant under rotation, i.e., when we consider a direction or another, we measure the same response;
- characterized by a *fading memory* when the material response depends on the very recent history.

The reader can refer to the book by Hutter & Jöhnk (2004) for further developments on *symmetry* in materials and their consequences in the constitutive equations.

1.2.5 Representation theorems

Cayley-Hamilton theorem

In linear algebra, the *Cayley-Hamilton theorem* says that from a second order tensor $\mathbf{M}$, we can build a third-order polynomial

$$P_M(\lambda) = \det(\mathbf{M} - \lambda\mathbf{1}) = -\lambda^3 + M_I\lambda^2 - M_{II}\lambda + M_{III},$$

where M_I, M_{II}, and M_{III} are called the fundamental invariants of $\mathbf{M}$, with

$$M_I = \operatorname{tr}\mathbf{M},\ M_{II} = \frac{1}{2}\left((\operatorname{tr}\mathbf{M})^2 - \operatorname{tr}\mathbf{M}^2\right), \text{ and } M_{III} = \det\mathbf{M}.$$

The zeros of this polynomial are the eigenvalues of $\mathbf{M}$ and, moreover, we have the relation

$$P_M(\mathbf{M}) = 0.$$

The remarkable result is that it is possible to define three independent scalar quantities that are objective, i.e., they do not depend on the frame in which we can express $\mathbf{M}$ since they

are scalar. From this viewpoint, it is equivalent to write $f = f(\mathbf{M})$ or $f = f(M_I, M_{II}, M_{III})$. The benefit is twofold

- it is mostly easier to handle scalar quantities than tensors;
- we can reduce the number of variables needed for describing the behavior. For instance, instead of a second-order symmetric tensor (with 6 independent variables), we can use the three invariants without loss of information.

We will now show that it is possible to interpret the invariants physically.

Physical interpretation

Any combination of invariants is invariant. Using this principle, we can build invariants that are physically meaningful. For the stress tensor $\boldsymbol{\sigma}$ and the strain tensor $\boldsymbol{\epsilon}$, we introduce

- the first invariants of the stress tensor and the strain tensor are defined as

$$I_{1,\sigma} = \operatorname{tr} \boldsymbol{\sigma} = 3\sigma_m \text{ and } I_{1,\epsilon} = \operatorname{tr} \boldsymbol{\epsilon} = \frac{\Delta V}{V},$$

  showing that the first invariant of the stress tensor gives an idea of the mean pressure at a given point, while the first invariant of the strain tensor specifies the relative volume variation.

- the second invariant is usually defined as

$$I_{2,\sigma} = \operatorname{tr} \mathbf{s}^2 \text{ and } I_{2,\epsilon} = \operatorname{tr} \mathbf{e}^2,$$

  where $\mathbf{s} = \boldsymbol{\sigma} - \sigma_m \mathbf{1}$ is the deviatoric stress tensor and $\mathbf{e} = \boldsymbol{\epsilon} - I_{1,\epsilon}\mathbf{1}/3$ is the deviatoric strain tensor. The second invariant indicated how large the departure from the mean state is.

- the third invariant is mostly defined as

$$I_{3,\sigma} = \operatorname{tr} \mathbf{s}^3 \text{ and } I_{3,\epsilon} = \operatorname{tr} \mathbf{e}^3.$$

  We can show that the third invariant makes it possible to define a phase angle in the deviatoric plane[3].

$$\cos\phi = -\frac{\sqrt{6} I_3}{I_2^{3/2}}.$$

To illustrate these notions, let us assume that we know the stress tensor at a given point M. The stress tensor being symmetric, we know that there there is an orthogonal basis made up of the eigenvectors of $\boldsymbol{\sigma}$. In the stress space, where the coordinates are given by the eigenvalues of $\boldsymbol{\sigma}$ (hereafter referred to as σ_i with $1 \leq i \leq 3$), the stress state at M is represented by a point.

The position of this stress point can be given in terms of the Cartesian coordinates or in terms of invariants. Indeed, let us call H the projection of M onto the first trissectrice (straight line $\sigma_1 = \sigma_2 = \sigma_3$). To locate M, we need to know the distances OH and HM

[3] plane passing through H and the normal of which is the first trissectrice

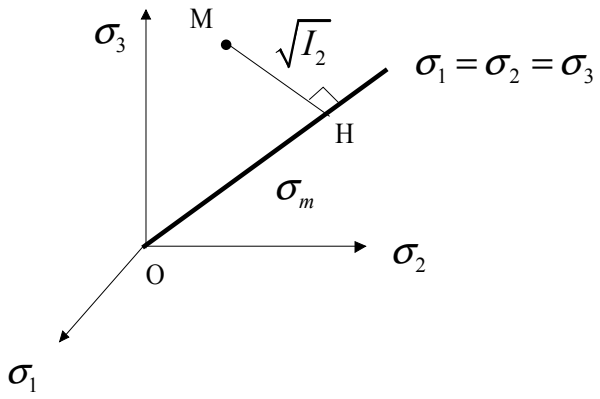


Figure 1.7. *Stress space.*

together with an angle ϕ with respect to an arbitrary direction in the deviatoric plane. It is straightforward to show that

- OH represents the mean stress at M;
- HM represents the departure from an isotropic state and $|HM|^2 = I_{2,\sigma} = s_1^2 + s_2^2 + s_3^2$ where s_i are the eigenvalues of the deviatoric stress tensor.

Representation theorems

Representation theorems are an ensemble of rules that specify how to transform a tensor-valued expression into an expression involving invariants (Boehler, 1987; Zheng, 1994).

For instance, let us assume that we have to compute the strain energy function W of an elastic body. We can write that $W = W(\mathbf{E})$, where $\mathbf{E}$ is the strain tensor, or we can write $W = W(E_i)$ with E_i with $1 \leq i \leq 6$ the six independent components of $\mathbf{E}$. Using invariants, we can also write

$$W = W(I_{1,\epsilon},\ I_{2,\epsilon},\ I_{3,\epsilon}).$$

In doing so, we reduce the number of variables from 6 to 3. If we assume that there is a preferred direction of deformation $\mathbf{n}$, then we can write $W = W(\mathbf{E},\ \mathbf{n})$ (i.e., 9 variables) or

$$W = W(I_{1,\epsilon},\ I_{2,\epsilon},\ I_{3,\epsilon},\ \mathbf{n}\cdot\mathbf{E}\cdot\mathbf{n},\ \mathbf{n}\cdot\mathbf{E}^2\cdot\mathbf{n}),$$

reducing the number of variables from 9 to 5.

Representation of tensor-valued functions in complete irreducible forms has been proved to be very helpful in formulating nonlinear constitutive equations for isotropic or anisotropic materials.

1.2.6 Balance equations

Transport theorem

In any course on functional analysis, one can find the Leibnitz formula that shows how to derive an integral real-valued function, the boundaries of which may vary with time

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{a(t)}^{b(t)} f(x,\ t)\mathrm{d}t = \int_{a(t)}^{b(t)} \frac{\partial f(x,\ t)}{\partial t}\mathrm{d}x + f(b(t))\frac{\mathrm{d}b}{\mathrm{d}t} - f(a(t))\frac{\mathrm{d}a}{\mathrm{d}t}. \tag{1.9}$$

Leibnitz relation can be generalized to multiple integrals, i.e. integration is made on volumes instead of intervals. One obtains the following relation called *transport theorem*

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\int_V f\mathrm{d}V = \int_V \frac{\partial f}{\partial t}\mathrm{d}V + \int_S f\mathbf{u}\cdot\mathbf{n}\mathrm{d}S,} \tag{1.10}$$

where V is the control volume containing a given mass of fluid, S is the surface bounding this volume, and $\mathbf{n}$ is the vector normal to the surface S ; $\mathbf{n}$ is unitary ($|\mathbf{n}| = 1$) and outwardly oriented. This relation written here for a scalar function f holds for any vectorial function $\mathbf{f}$.

Equation (1.10) is fundamental since it makes it possible to derive all the equations needed in continuum mechanics. It can be interpreted as follows. Any variation in f over time within the control volume V results from

- local change in f with time;
- flux of f through S (flux = inflow – outflow through V).

An helpful variant of the transport theorem is obtained by using the Green-Ostrogradski theorem :

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\int_V f\mathrm{d}V = \int_V \left(\frac{\partial f}{\partial t} + \nabla\cdot(f\mathbf{u})\right)\mathrm{d}V}.$$

This expression shall be used to derive local governing equations in the following.

A control volume is most often a material volume, i.e., it is made up of a collection of particles that are followed up in their motion; its borders are fluid and move with the fluid, which means that the boundary velocity corresponds to the local velocity at the boundary. On some occasions, we can also define an arbitrary control volume, the velocity of which $\mathbf{u}$ at the border surface S does not correspond to that of the fluid.

Another important variant of the transport theorem is the *Reynolds theorem* that applies to integrands that take the form ρf, with ρ the fluid density. This theorem reads:

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho f\mathrm{d}V = \int_V \rho\frac{\mathrm{d}}{\mathrm{d}t}f\mathrm{d}V.} \tag{1.11}$$

Conservation of mass

Let us apply the transport theorem (1.10) to $f = \rho$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho\mathrm{d}V = \int_V \frac{\partial\rho(\mathbf{x},\ t)}{\partial t}\mathrm{d}V + \int_S \rho\mathbf{u}\cdot\mathbf{n}\mathrm{d}S.$$

Making use of the divergence theorem (Green-Ostrogradski), we find

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho \mathrm{d}V = \int_V \left(\frac{\partial \rho(\mathbf{x},\ t)}{\partial t} + \nabla \cdot (\rho \mathbf{u})\right) \mathrm{d}V = 0.$$

Note that the total mass is constant because we follow up a finite number of particles and there is no production or loss of particles.

When ρ is continuous, then we can pass from a control-volume formulation to a local equation

$$\boxed{\frac{\partial \rho(\mathbf{x},\ t)}{\partial t} + \nabla \cdot \rho \mathbf{u} = 0.} \tag{1.12}$$

This equation is often called *continuity equation*. It can also be cast in the following form:

$$\frac{1}{\rho}\frac{\mathrm{d}\rho}{\mathrm{d}t} = -\rho \nabla \cdot \mathbf{u}.$$

Recall that passing from global to local equations is permitted only if the field is continuous. This is not always the case, e.g. when there is a *shock* inside the control volume. Specific equation must be used (see below).

Here are other helpful definitions

- a flow is said *isochoric* when $\frac{1}{\rho}\frac{\mathrm{d}\rho}{\mathrm{d}t} = 0$ (e.g., when the Mach number is less than unity, air flow is isochoric);
- a material is said *incompressible* when ρ is constant at any point and any time (water can be considered as incompressible under normal flow conditions).

Conservation of momentum

One applies the transport theorem (1.10) tho the momentum $f = \rho \mathbf{u}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho \mathbf{u} \mathrm{d}V = \int_V \frac{\partial \rho \mathbf{u}}{\partial t} \mathrm{d}V + \int_S \rho \mathbf{u}(\mathbf{u} \cdot \mathbf{n}) \mathrm{d}S.$$

There are many variants of this equation, based either on different ways of expressing the material derivative of $\rho \mathbf{u}$ or on different ways of expressing the velocity (e.g., streamline function, vorticity). On applying the divergence theorem, one gets :

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho \mathbf{u} \mathrm{d}V = \int_V \left(\frac{\partial \rho \mathbf{u}}{\partial t} + \nabla \cdot \rho \mathbf{u}\mathbf{u}\right) \mathrm{d}V,$$

or equivalently

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho \mathbf{u} \mathrm{d}V = \int_V \rho \left(\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathbf{u}\mathbf{u}\right) \mathrm{d}V.$$

The fundamental principle of Mechanics is that any time variation in momentum results from applying body or surface force(s) on the control volume

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho \mathbf{u} \mathrm{d}V = \text{forces applied on } V.$$

Once again, we can obtain a local expression of the momentum balance equation if the fields are continuous:

$$\frac{\partial \rho \mathbf{u}}{\partial t} + \nabla \cdot \rho \mathbf{u}\mathbf{u} = \rho \mathbf{g} + \nabla \cdot \boldsymbol{\sigma},$$

or

$$\boxed{\rho \frac{\partial \mathbf{u}}{\partial t} + \rho \mathbf{u} \cdot \nabla \mathbf{u} = \rho \mathbf{g} + \nabla \cdot \boldsymbol{\sigma} = \rho \mathbf{g} - \nabla p + \nabla \cdot \mathbf{s}} \tag{1.13}$$

Be aware that $\mathbf{u} \cdot \nabla \mathbf{u}$ does not mean the product between $\mathbf{u}$ and the tensor $\nabla \mathbf{u}$. Rigorously speaking, it would be better to write $(\mathbf{u} \cdot \nabla)\mathbf{u}$, the parentheses are used to say that $\mathbf{u} \cdot \nabla$ is a differential operator applied to $\mathbf{u}$.

1.2.7 Conservation of energy

Kinetic energy

The transport theorem (1.10) is now applied to the kinetic energy $f = E_c = \frac{1}{2}\rho|\mathbf{u}|^2$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \frac{1}{2}\rho|\mathbf{u}|^2 \mathrm{d}V = \int_V \frac{\partial E_c}{\partial t}\mathrm{d}V + \int_S \frac{1}{2}\rho|\mathbf{u}|^2(\mathbf{u}\cdot\mathbf{n})\mathrm{d}S.$$

As earlier for the momentum equation, there are variants of this equation depending on how the material derivative of E_c is expressed. Making use of the divergence theorem leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V E_c \mathrm{d}V = \int_V \left(\frac{\partial E_c}{\partial t} + \nabla \cdot \left(\frac{1}{2}\rho|\mathbf{u}|^2\mathbf{u}\right)\right)\mathrm{d}V,$$

Note that the equation can be inferred from (1.13) by multiplying it by $\mathbf{u}$, then replacing terms such as $\mathbf{u}\partial\mathbf{u}$ with $\partial|\mathbf{u}|^2/2$, and then integrating over the control volume V. We then deduce the bulk expression of the *kinetic energy theorem*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V E_c \mathrm{d}V = \text{power supplied to the volume } V - \text{power dissipated in } V,$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V E_c \mathrm{d}V = \int_V \rho \mathbf{u}\cdot\mathbf{g}\mathrm{d}V + \int_V \mathbf{u}\cdot(\nabla\cdot\boldsymbol{\sigma})\mathrm{d}V.$$

When the fields are continuous and making use of

$$\mathbf{u}\cdot(\nabla\cdot\boldsymbol{\sigma}) = \nabla\cdot(\mathbf{u}\cdot\boldsymbol{\sigma}) - \boldsymbol{\sigma} : \nabla\mathbf{u},$$

we can derive the local equation

$$\boxed{\frac{\partial E_c}{\partial t} + \nabla\cdot(E_c\mathbf{u}) = \rho\mathbf{u}\cdot\mathbf{g} + \nabla\cdot(\mathbf{u}\cdot\boldsymbol{\sigma}) - \boldsymbol{\sigma} : \nabla\mathbf{u}.} \tag{1.14}$$

We refer to $\Phi = \boldsymbol{\sigma} : \nabla\mathbf{u}$ as the *energy dissipation rate*. Using the decomposition of the strain rate tensor into its symmetric and antisymmetric parts, we also obtain

$$\Phi = \boldsymbol{\sigma} : \mathbf{d} = \mathrm{tr}\,(\boldsymbol{\sigma}\cdot\mathbf{d}).$$

Recall that the energy balance theorem contains nothing more that the momentum balance

equation does. For a regular problem, we can select either theorem; the choice is a matter of personal convenience or strategy (for alleviating computation). On some occasions, only one of these equations can be used in practice. For instance, when studying shock formation, it is usually better to use momentum balance equations because shocks induce energy dissipation that is not easy to compute.

$\Rrightarrow$ In many practical applications (incompressible fluid in a steady regime), the energy balance equation can be transformed into the Bernoulli equation

$$\boxed{\frac{\partial E_c}{\partial t} + \nabla \cdot \left(\mathbf{u} \frac{\rho |\mathbf{u}|^2 + 2p_*}{2} \right) = \nabla \cdot (\mathbf{u} \cdot \mathbf{s}) - \mathbf{s} : \mathbf{d},}$$

where $p_* = p + \psi$ is the *generalized pressure*, with $\psi = \rho g z$ the gravity potential ($\rho \mathbf{g} = -\nabla \psi$). The *Bernoulli theorem* is obtained by assuming that

- the flow is steady, i.e. $\partial E_c / \partial t = 0$;
- the Reynolds number is high (the fluid is inviscid).

We then obtain

$$\nabla \cdot \left(\mathbf{u} \frac{\rho |\mathbf{u}|^2 + 2p_*}{2} \right) = \mathbf{u} \cdot \nabla \left(\frac{\rho |\mathbf{u}|^2}{2} + p_* \right) = 0,$$

which means that

$$\boxed{\Psi = \frac{1}{2} \rho |\mathbf{u}|^2 + p_* = \frac{1}{2} \rho |\mathbf{u}|^2 + p + \rho g z} \tag{1.15}$$

is constant along a streamline.

First axiom of thermodynamics

The first law of thermodynamics applied to a control volume V is the following statement

$$\begin{aligned} \text{rate of change of total energy} = \ & \text{rate of change of work of the forces applied to } V \\ & + \text{ rate of heat addition.} \end{aligned} \tag{1.16}$$

We define the following quantities

- The *total energy* is the sum of the internal energy $E = \rho e$ (e internal energy per unit mass) and the kinetic energy $\rho u^2/2$.
- The rate of work is the power of the forces applied to the boundary and the body forces (e.g., gravity).
- The heat supplied to the control volume results from heat generated at points within V and from heat transmitted through the boundaries. The heat supply density is denoted by r and the heat flux is written $q = \mathbf{q} \cdot \mathbf{n}$, with $\mathbf{q}$ the heat flux vector (Stokes relation). When the Fourier law holds, the heat flux vector is related to the gradient temperature $\mathbf{q} = -\kappa \nabla T$, where κ is the *heat conductivity*.

Translated into mathematical terms, this statement becomes

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho\left(e+\frac{1}{2}u^2\right)\mathrm{d}V = \int_V \rho\mathbf{u}\cdot\mathbf{g}\mathrm{d}V + \int_A \mathbf{u}\cdot(\boldsymbol{\sigma}\cdot\mathbf{n})\mathrm{d}A + \int_V \rho r\mathrm{d}V - \int_A \mathbf{q}\cdot\mathbf{n}\mathrm{d}A,}$$

where A is the surface bounding V. The local form is

$$\rho\frac{\mathrm{d}}{\mathrm{d}t}\left(e+\frac{1}{2}u^2\right) = \rho\mathbf{u}\cdot\mathbf{g} + \nabla\cdot(\boldsymbol{\sigma}\mathbf{u}) - \nabla\mathbf{q} + \rho r.$$

Making use of the kinetic energy theorem leads to the following relation

$$\boxed{\rho\frac{\mathrm{d}e}{\mathrm{d}t} = \Phi - \nabla\mathbf{q} + \rho r,}$$

where $\Phi = \boldsymbol{\sigma} : \mathbf{d}$ is the dissipation rate (also called stress power). Variations in the internal energy are caused by (viscous) dissipation, the flux of heat, and/or a source/sink of heat.

Second axiom of thermodynamics

We end our discussion on energy conservation by recalling the second principle used in thermodynamics, which reflects the irreversibility of time processes associated with energy dissipation. The second law states that

$$\rho\dot{f} - \Phi + \rho\dot{T}s + \frac{1}{T}\mathbf{q}\cdot\nabla T \leq 0$$

with $f = e - Ts$ the free energy per unit mass, s the entropy per unit mass, Φ the energy dissipation rate, $\mathbf{q}$ the heat flux vector, and T the temperature.

1.2.8 Jump conditions

In practice, there a number of situations where there are rapid changes in the flow features over relatively short distances. For instance, a high-speed airplane or spatial capsule creates a shock wave as it breaks the sound barrier, as is shown in Figures 1.8 and 1.9

In theory, it is usually appropriate to consider the shock as a discontinuity surface, i.e. a surface through which some of the flow variables (density, velocity, etc.) may be become discontinuous. The local balance equations are valid on either side of the jump, but not at the shock surface. This implies that we have to specify the jump conditions on the flow variables induced by a shock. Note that a discontinuity surface may be an existing boundary (e.g. a free surface) or it may be created under some flow conditions. Spontaneous creation or disappearance of a shock surface is typical for hyperbolic partial differential equations (Courant & Friedrich, 1948). We will focus here on the latter case (see also § 3.2.2).

Let us first consider the scalar case. We have to solve a hyperbolic problem in the form:

$$\frac{\partial f(x,\ t)}{\partial t} + \frac{\partial G[f(x,\ t)]}{\partial x} = a(x,\ t),$$

where G is a function and a another function called the source term. Note that the equations looks like the balance equations we have seen earlier. Usually such an equation originates

Figure 1.8. *A U.S. Navy airplane creates a shock wave as it breaks the sound barrier. The shock wave is visible as a large cloud of condensation formed by the cooling of the air. A smaller shock wave can be seen forming on top of the canopy.(U.S. Navy photo by John Gay).*

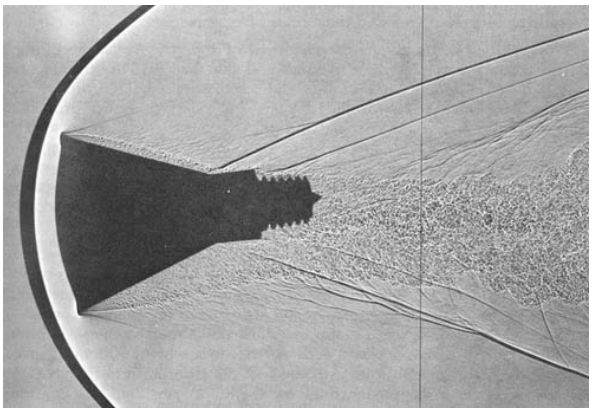


Figure 1.9. *A shadowgraph of the Project Mercury reentry capsule, showing the bow-shock wave in front of it and the flow fields behind the capsule. Photograph from NASA).*

from the conservation of a quantity in a given control volume, i.e. an equation in the following integral form

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V f(x,\ t)\mathrm{d}x = G[f(x_2,\ t)] - G[f(x_1,\ t)] + \int_V a(x,\ t)\mathrm{d}x,$$

where the control volume corresponds to the range $[x_1,\ x_2]$ in the scalar case. When f is continuous over V, the two equations are equivalent. Let us assume that there is a moving point $x = s(t)$ within $[x_1,\ x_2]$ at which f admits a discontinuity. Making use of the Leibnitz rule, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{x_1}^{x_2} f(x,\ t)\mathrm{d}x = \int_{x_1}^{s(t)} \frac{\partial f(x,\ t)}{\partial t}\mathrm{d}x + \int_{s(t)}^{x_2} \frac{\partial f(x,\ t)}{\partial t}\mathrm{d}x - \dot{s}[\![f]\!],$$

where we have broken down $[x_1,\ x_2] = [x_1,\ s(t)] + [s,\ x_2(t)]$ and where $[\![f]\!]$ is the jump experienced by f:

$$[\![f]\!] = \lim_{x\to s,\ x>0} f(x) - \lim_{x\to s,\ x<0} f(x).$$

Then taking the limits $x_1 \to s$ and $x_2 \to s$, leads to

$$[\![G[f(x,\ t)] - \dot{s} f(x,\ t)]\!] = 0.$$

The quantity is $G[f] - f\dot{s}$ is conserved through the shock. We can also deduce the shock velocity

$$\dot{s} = \frac{[\![G[f(x,\ t)]]\!]}{[\![f(x,\ t)]\!]}.$$

For instance, we can retrieve the *Rankine-Hugoniot shock conditions* used in gas dynamics if we take

- $f = \rho u$ (u being a velocity and ρ a density), $G[f] = \frac{1}{2}\rho u^2 + p$, and $a = 0$ (with p the pressure), we have the (scalar) momentum equation along an axis. The shock condition is then $[\![p + \frac{1}{2}\rho u^2]\!] = \dot{s}[\![\rho u]\!]$ at $x = s(t)$.
- $f = \rho$, $G[f] = \rho u$, and $a = 0$, we have the (scalar) mass equation along an axis. The shock condition is then $[\![\rho u]\!] = \dot{s}[\![\rho]\!]$ at $x = s(t)$. If we introduce the relative velocity $u' = u - \dot{s}$, we have also $[\![\rho u']\!] = 0$, which means that in the frame relative to the shock, the mass flux is conserved.

This equation can be generalized to higher dimensions without any problem.

When dealing with shocks, it is very important to use the original conservation equations (in an integral form) from which the local equation has been derived. Typically, when the field are continuous, the equations

$$\frac{\partial \rho u}{\partial t} + u\frac{\partial \rho u}{\partial x} = 0,$$

and

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0,$$

are equivalent because of the continuity equation. However, the former equation comes directly from the conservation of momentum (hence it is called the *conservative* form and ρu a conservative variable), whereas the latter is a simplification (called *non-conservative* form). It we reintegrate these equations to obtain the shock conditions, we will not find the same shock velocity. For this reason, care must be taken in computing shock conditions related to a hyperbolic partial differential equation.

1.3 Phenomenological constitutive equations

In many cases, most of the available information on the rheological behavior of a material is inferred from simple shear experiments. But, contrary to the Newtonian (linear) case, the tensorial form cannot be merely and easily generalized from the scalar expression fitted to experimental data. Earlier in this chapter, we have seen that:

- First, building a three-dimensional expression of the stress tensor involves respecting a certain number of formulation principles. These principles simply express the idea that the material properties of a fluid should be independent of the observer or frame of reference (principle of material objectivity) and the behavior of a material element depends only on the previous history of that element and not on the state of neighboring elements (Bird *et al.*, 1987).
- Then it is often necessary to provide extra information or rules to build a convenient expression for the constitutive equation.

We shall illustrate this with several examples.

1.3.1 Newtonian behavior

We start with an application of the representation theorem and the virtual power principle. We have an isotropic, incompressible, homogenous material assumed to be viscous. When it deforms, the energy dissipation rate Φ is fonction of the state of deformation, or more specifically of the strain rate $\mathbf{d}$, hence we write

$$\Phi = \Phi(\mathbf{d}),$$

but, making use of the representation, we can directly deduce the equivalent but more reduced form

$$\Phi = \Phi(d_I,\ d_{II},\ d_{III}),$$

where d_i represents the ith invariant of the strain-rate tensor. Since the material is isotropic, Φ does not depend on d_{III}; since it is incompressible, we have $d_I = 0$. Now we assume that we have a linear behavior, i.e., the energy rate dissipation must be a quadratic function of the invariants since dissipation = stress $\times$ strain rate $\propto$ strain rate^2. So we get

$$\Phi = \Phi(d_I,\ d_{II},\ d_{III}) = -\alpha d_{II},$$

with $\alpha > 0$ a constant and $d_{II} = -\frac{1}{2}\mathrm{tr}\,\mathbf{d}^2 = -\frac{1}{2}d_{kl}d_{kl}$ (Einstein's convention used). Using (1.8), we show that the extra-stress tensor is defined as

$$s_{ij} = \frac{\partial \Phi}{\partial d_{ij}} = \alpha\frac{1}{2}\frac{\partial d_{kl}d_{kl}}{\partial d_{ij}} = \alpha d_{kl}\delta_{ik}\delta jl = \alpha d_{ij},$$

from which we retrieve the Newtonian constitutive equation by posing $\alpha = 2\mu$.

1.3.2 Viscoplastic behavior

When a fluid exhibits viscoplastic properties, we usually fit experimental data with a Bingham equation as a first approximation (Bird *et al.*, 1983):

$$\dot{\gamma} > 0 \Rightarrow \tau = \tau_c + K\dot{\gamma} \tag{1.17}$$

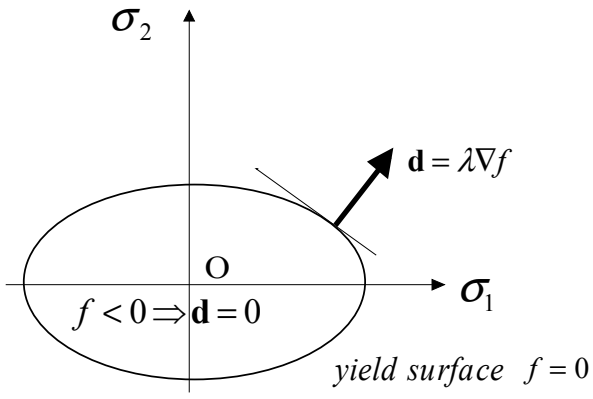


Figure 1.10. *Yield surface delimiting two domains.*

Equation (1.17) means that

- for shear stresses in excess of a critical value, called the yield stress, the shear stress is a linear function of the shear rate;
- conversely when $\tau \leq \tau_c$ there is no shear within the fluid ($\dot{\gamma} = 0$).

The question arises as to how the scalar expression can be transformed into a tensorial form. The usual way (but not the only one) is to consider a process, called *plastic rule*, as the key process of yielding.

A plastic rule includes two ingredients.

- First, it postulates the existence of a surface in the stress space $(\sigma_1, \sigma_2, \sigma_3)$ delimiting two possible mechanical states of a material element (σ_i denotes a principal stress, that is an eigenvalue of the stress tensor), as depicted in Fig. 1.10. The surface is referred to as the *yield surface* and is usually represented by an equation in the form $f(\sigma_1, \sigma_2, \sigma_3) = 0$. When $f < 0$, behavior is generally assumed to be elastic or rigid. When $f = 0$, the material yields.
- Second it is assumed that, after yielding, the strain-rate is directly proportional to the surplus of stress, that is, the distance between the point the representing the stress state and the yield surface. Translated into mathematical terms, this leads to write: $\dot{\gamma} = \lambda \nabla f$ with λ a proportionality coefficient (Lagrangian multiplier).

How must the yield function f be built to satisfy the principle of material objectivity? For f to be independent of the frame, it must be expressed not as a function of the components of the stress tensor, but as a function of its invariants:

- the first invariant $I_1 = \mathrm{tr}\,\boldsymbol{\sigma} = \sigma_1 + \sigma_2 + \sigma_3$ represents the *mean stress* multiplied by 3 ($|\mathbf{OP}| = I_1/3$ in Fig. 1.11);
- the second invariant $I_2 = (\mathrm{tr}^2\boldsymbol{\sigma} - \mathrm{tr}\boldsymbol{\sigma}^2)/2 = -\mathrm{tr}(\mathbf{s}^2)/2$ can be interpreted as the deviation of a stress state from the mean stress state ($|\mathbf{PM}|^2 = -2I_2$ in Fig. 1.11) and is accordingly called the *stress deviator*;

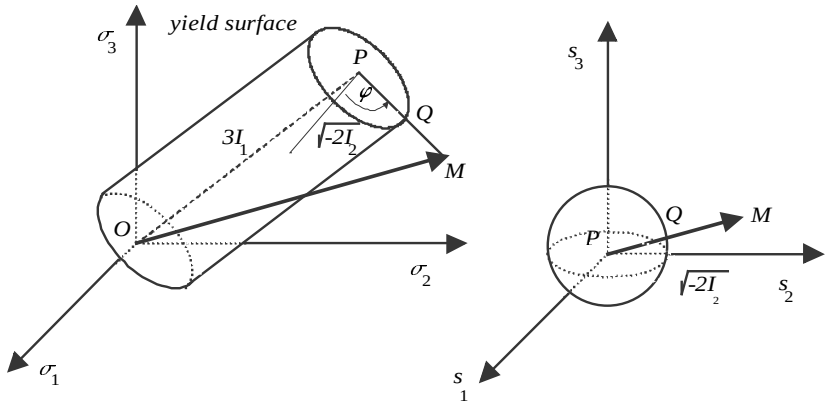


Figure 1.11. *On the left, the yield surface in the stress space when the von Mises criterion is selected as yield function. A stress state is characterized by its three principal stresses and thus can be reported in the stress space. The three invariants of the stress tensor can be interpreted in terms of coordinates.*

- the third invariant $I_3 = -\mathrm{tr}\,\mathbf{s}^3/6$ reflects the angle in the deviatoric plane made by the direction **PM** with respect to the projection of -axis and is sometimes called the *phase* ($\cos^2 3\varphi = I_3^2/I_2^3$ in Fig. 1.11).

If the material is isotropic and homogenous, the yield function f is expected to be independent of the mean pressure and the third invariant. Thus we have $f(\sigma_1, \sigma_2, \sigma_3) = f(I_2)$. In plasticity, the simplest yield criterion is the *von Mises criterion*, asserting that yield occurs whenever the deviator exceeds a critical value (whose root gives the yield stress): $f(I_2) = \sqrt{-I_2} - \tau_c$. As depicted in Fig. 1.11, the yield surface is a cylinder of radius τ_c centered around an axis $\sigma_1 = \sigma_2 = \sigma_3$. If we draw the yield surface in the extra-stress space, we obtain a sphere of radius $\sqrt{2}\tau_c$.

Once the stress state is outside the cylinder defined by the yield surface, a flow occurs within the material. In a linear theory it is further assumed that the strain rate is proportional to the increment in stress (distance from the point representing the stress state M to the yield surface) and collinear to the normal $\partial f/\partial \mathbf{s}$. This leads to the expression:

$$\mathbf{d} = \lambda \left(\sqrt{I_2} - \tau_c\right) \frac{\mathbf{s}}{\sqrt{I_2}} \tag{1.18}$$

For convenience, we define the proportionality coefficient as: $\lambda^{-1} = 2\eta$. It is generally more usual to express the constitutive equation in the converse form. First, the second invariant of the strain rate tensor J_2 can be expressed as $J_2 = -\mathrm{tr}(\mathbf{d}^2)/2 = \left(\lambda\left(\sqrt{-I_2} - \tau_c\right)\right)^2$. Then we deduce the usual form of the Bingham constitutive equation:

$$\mathbf{d} = 0 \Leftrightarrow \sqrt{-I_2} \leq \tau_c \tag{1.19}$$

$$\mathbf{d} \neq 0 \Leftrightarrow \boldsymbol{\sigma} = -p\mathbf{1} + \left(2\eta + \frac{\tau_c}{\sqrt{-J_2}}\right)\mathbf{d} \tag{1.20}$$

It is worth noting that contrary to the Newtonian case, the general tensorial expression (1.19) cannot not easily be extrapolated from the steady simple-shear equation (1.17).

1.3.3 Viscoelasticity

When studying the linear 1D Maxwell model in § 5.5.4, we obtained an equation in the form

$$\frac{\mathrm{d}\gamma}{\mathrm{d}t} = \frac{1}{G}\frac{\mathrm{d}\tau}{\mathrm{d}t} + \frac{\tau}{\mu},$$

where G is the elastic modulus and μ denotes viscosity. We would like to transform this empirical equation into a 3D tensorial expression and naively we write

$$2\mu\mathbf{d} = \frac{\mu}{G}\frac{\mathrm{d}\boldsymbol{\sigma}}{\mathrm{d}t} + \boldsymbol{\sigma},$$

but it is not difficult to see that this expression does not satisfy the objectivity principle. According to this principle, the stress tensor does not depend on the frame in which we use it (or its components) or, in other words, it must be invariant under any rotation. Let us consider the stress tensor when the frame of reference is rotated. We have

$$\boldsymbol{\sigma}' = \mathbf{R}\cdot\boldsymbol{\sigma}\cdot\mathbf{R}^\dagger$$

the image of $\boldsymbol{\sigma}$, where $\mathbf{R}$ is an orthogonal tensor. Taking the time derivative leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\sigma}' = \frac{\mathrm{d}\mathbf{R}}{\mathrm{d}t}\cdot\boldsymbol{\sigma}\cdot\mathbf{R}^\dagger + \mathbf{R}\cdot\frac{\mathrm{d}\boldsymbol{\sigma}}{\mathrm{d}t}\cdot\mathbf{R}^\dagger + \mathbf{R}\cdot\boldsymbol{\sigma}\cdot\frac{\mathrm{d}\mathbf{R}^\dagger}{\mathrm{d}t} \neq \mathbf{R}\cdot\frac{\mathrm{d}\boldsymbol{\sigma}}{\mathrm{d}t}\cdot\mathbf{R}^\dagger,$$

which shows that $\dot{\boldsymbol{\sigma}}$ is not objective. To overcome this issue, we have to define a kind of time derivative that satisfies the objectivity principle. For this purpose, note that if we replace $\boldsymbol{\sigma}$ with $\mathbf{R}^\dagger\cdot\boldsymbol{\sigma}'\cdot\mathbf{R}$ and introduce $\boldsymbol{\Omega} = \dot{\mathbf{R}}\cdot\mathbf{R}^\dagger$, we have $\dot{\boldsymbol{\sigma}}' = \mathbf{R}\cdot\boldsymbol{\sigma}\cdot\mathbf{R} + \boldsymbol{\Omega}\cdot\boldsymbol{\sigma}' + \boldsymbol{\sigma}'\cdot\boldsymbol{\Omega}^\dagger$, which can be transformed by making use of the antisymmetry of $\boldsymbol{\Omega}$: $\dot{\boldsymbol{\sigma}}' = \mathbf{R}\cdot\dot{\boldsymbol{\sigma}}\cdot\mathbf{R}^\dagger + \boldsymbol{\Omega}\cdot\boldsymbol{\sigma}' - \boldsymbol{\sigma}'\cdot\boldsymbol{\Omega}$. If we want to transform this derivative into an objective derivative, we have to get rid of the last two contributing terms. There are different possibilities. Oldroyd introduced the Oldroyd (or convective contravariant) derivative as follows:

$$\overset{\triangle}{\boldsymbol{\sigma}} = \dot{\boldsymbol{\sigma}} - \mathbf{L}\cdot\boldsymbol{\sigma} - \boldsymbol{\sigma}\cdot\mathbf{L}^\dagger,$$

where $\mathbf{L}$ is the velocity gradient tensor. In this way, it is possible to

- provide a proper tensorial formulation of the constitutive equation;
- make an empirical law (primarily valid for small deformations) consistent with large deformations.

The stress tensor is then solution to

$$2\mu\mathbf{d} = \frac{\mu}{G}\overset{\triangle}{\mathbf{s}} + \mathbf{s}.$$

Note that in this particular case, we do not have necessarily $\operatorname{tr}\mathbf{s}$, which is a typical example of a fluid for which the extra-stress tensor differs from the deviatoric the stress tensor.

Exercises

Exercise 1.1 Let $\mathbf{S}$ and $\mathbf{A}$ symmetric and antisymmetric tensors. Show that

$$\mathrm{tr}\,(\mathbf{S}\cdot\mathbf{A}) = 0.$$

Exercise 1.2 Give the definition of a streamline. Show that an equation in a Cartesian frame is

$$\frac{\mathrm{d}x}{u} = \frac{\mathrm{d}y}{v} = \frac{\mathrm{d}z}{w},$$

where $(u,\ v,\ w)$ are the components of the velocity field.

Exercise 1.3 We consider a velocity field of the following form:

$$\mathbf{u} = \mathbf{\Gamma}\cdot\mathbf{x},$$

with

$$\mathbf{\Gamma} = \frac{1}{2}\dot{\gamma}\begin{bmatrix} 1+\alpha & 1-\alpha & 0 \\ -1+\alpha & -1-\alpha & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$\dot{\gamma}$ denotes the shear rate, and α is a constant satisfying $0 \leq \alpha \leq 1$. Characterize the velocity field depending on α. Hint: compute $\nabla\mathbf{u}$, $\mathbf{D}$ and $\mathbf{W}$. Is flow isochoric?

Exercise 1.4 Let us consider the following velocity field in the upper plane $y > 0$:

$$u = kx \text{ and } v = -ky,$$

with $k > 0$ and where $(u,\ v)$ are the velocity components. Flow is steady. The lower plane $y \leq 0$ is a solid body; the interface at $y = 0$ is then solid.

- compute the tensor "velocity gradient" associated with this velocity field;
- deduce the "rate of strain" tensor;
- compute the stress state for point A (0, 1) ;
- compute the streamlines;
- are the usual boundary conditions satisfied? Are these those of a Newtonian fluid?

Exercise 1.5 Let us consider a steady uniform flow down an infinite plane, which is inclined at θ to the horizontal. The fluid is incompressible with density ϱ. Show that independently of the constitutive equation, the shear stress τ is

$$\tau(y) = \varrho g(h-y),$$

where h is the flow depth, y is the coordinate normal to the plane, and g is gravity acceleration.

We now consider the flow of a Newtonian fluid with viscosity μ. Compute the velocity profile for a steady uniform flow. What are the principal directions of the stress and strain-rate tensors?

2 Equations of mechanics

In this chapter we will take a look at different families of differential equations that we can meet when studying mechanical processes. Bearing in mind the different types of equations and physical phenomena involved will be important later in this course to understand the strategies used for solving practical problems.

2.1 Equation classification

2.1.1 Scalar equation

An equation is said to be *scalar* if it involves only scalar quantities, without differential term. In mechanics, most problems are differential and meeting with purely scalar equations is seldom. A notable exception is the Bernoulli equation which states that the quantity

$$\psi = \varrho\frac{u^2}{2} + \varrho g z + p$$

is constant under some flow conditions, with u fluid velocity, ϱ fluid density, p its pressure, g gravity acceleration, and z elevation with respect to a reference level.

2.1.2 Ordinary differential equation

An *ordinary differential equation* (ODE) is a differential equation where the function is differentiated with respect to a single variable (the independent variable). The ordinary differential equations are quite common:

- either because the problem is basically a one-dimensional problem;
- or because with the help of transformations, we can reduce a problem of partial differential equations (PDEs) to an ordinary differential problem, which is much easier to solve analytically or numerically.

♣ **Example**. – The Pascal equation in fluid statics is an ordinary differential equation

$$\frac{\mathrm{d}p}{\mathrm{d}z} + \varrho g = 0,$$

where ϱ denotes fluid density, p its pressure, g gravity acceleration, and z elevation with respect to a reference level.

The order of an ordinary differential equation is defined as the order of the highest derivative. The order determines the number of initial conditions that are needed to solve the differential equation.

♣ **Example**. – A differential equation of order 2 such that $y'' + ay' + by = c$ requires to specify two boundary conditions. These can be given at a point (for example, we may pose $y(0) = 0$ and $y'(0) = 1$) or at different points (for example, we may ask $y(0) = 0$ and $y'(1) = 1$). In the first case, it is called *initial-value problem*, whereas in the latter case, we refer to is as a *boundary-value problem*[1]. □

An ordinary differential equation is called *linear* if it involves only linear combinations of derivatives of the function and the function itself. For example, $x^3y'' + y' = 0$ is linear (in y), but $y'y'' + x^3 = 0$ is not linear. An equation is called *quasi-linear* if it consists of a linear combination of derivatives, but not necessarily of the function. For example, $yy' + x^2y = 1$ is not linear but quasi-linear.

An ordinary differential equation quasi-linear first order can be cast in the form

$$\frac{\mathrm{d}u}{\mathrm{d}x} = \frac{f(u,\ x)}{g(u,\ x)},$$

with f and g two functions of u and x. This is form is particularly helpful since it allows to obtaining graphical representation of the solution (phase portrait) and analysis of singular points. This equation can also be put in the following differential form

$$g(u,\ x)\mathrm{d}u - f(u,\ x)\mathrm{d}x = 0.$$

The latter form is used to find exact differentials, i.e., functions ψ such that $\mathrm{d}\psi = g(u,\ x)\mathrm{d}u - f(u,\ x)\mathrm{d}x$. If we are successful, we can obtain an implicit solution to the differential equation in the form $\psi(u,\ x) = const$.

2.1.3 Partial differential equations

Most fundamental equations of mechanics such as the Navier-Stokes equations are partial differential equations, i.e. they describe how the process varies—depending on time and location—by relating temporal and spatial derivatives. There is a wide variety of problems for PDEs we will uncover in what follows.

There are several ways to write a partial differential equation. For example, the diffusion equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

can be written with the short-hand notation $u_t = u_{xx}$ or $\partial_t u = \partial_{xx} u$.

[1] We distinguish two types of conditions because from a numerical point of view, we have to use different techniques depending on the conditions. When conditions are given at different points, we must employ such as "shooting methods" to solve the equations numerically.

Terminology

The *order* of a partial differential equation is the order of the higher differential term. For example, the equation $u_t = u_{xx}$ is order 2. The dependent variable is the function that we differentiate with respect to the independent variables; in the example above, u is the dependent variable, while x and t are the independent variables. The number of independent variables are the *dimension* of the partial differential equation. As for an ordinary differential equation, a partial differential equation is linear if it is linear in the dependent variable, the equation $u_t = u_{xx}$ is a linear equation because it depends linearly on u or its derivatives.

Classification of linear du second-order partial differential equations

The only general classification of partial differential equations concerns linear equations of second order. These equations are of the following form

$$au_{xx} + 2bu_{xy} + cu_{yy} + du_x + eu_y + fu = g, \tag{2.1}$$

where a, b, c, d, e, f, and g are real-valued functions x and y. When $g = 0$, the equation is said to be *homogeneous*. Linear equations are classified depending on the sign of $\Delta = b^2 - ac > 0$:

- if $\Delta = b^2 - ac > 0$, Equation (2.1) is *hyperbolic*. The wave equation (2.22) is an example. In fluid mechanics, transport equations are often hyperbolic. The *canonical* form is

  $$u_{xx} - u_{yy} + \cdots = 0 \text{ or, equivalently, } u_{xy} + \cdots = 0,$$

  where dots represent terms related to u or its first-order derivatives;

- if $\Delta = b^2 - ac < 0$, Equation (2.1) is *elliptic*. The Laplace equation (2.24) is an example. Equations describing equilibrium of a process are often elliptic. The canonical form is

  $$u_{xx} + u_{yy} + \cdots = 0$$

- if $\Delta = b^2 - ac = 0$, Equation (2.1) is *parabolic*. The heat equation (2.7) is an example. Diffusion equations are often parabolic. The canonical form is

  $$u_{yy} + \cdots = 0.$$

There is a strong link between the name given to differential equations and the name of conics. Indeed, if we assume that the coefficients of equation (2.1) are constant and we substitute into Equation(2.1) u_{xx} with x^2, u_x with x, u_{yy} with y^2, u_y with y, and u_{xy} with xy, we obtain the general equation of a conic, which depending on the sign $\Delta = b^2 - ac$ gives a parabola ($\Delta = 0$), an ellipse ($\Delta < 0$), or a hyperbola ($\Delta > 0$), as shown in Figure 2.1. This figure shows that the differential terms are linked and vary according to the constraints intrinsic to each type of curve. We note for example that for hyperbolic equations, there are two branches and part of the $x-y$ plane is not crossed by the curve, which allows discontinuous jumps from one branch to another; such jumps exist in differential equations and are called *shocks*: a hyperbolic equation is able to generate solutions that become discontinuous, i.e. undergo a shock even if initially they were continuous.

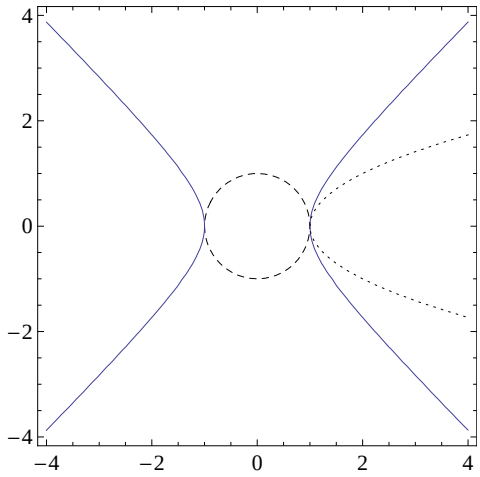


Figure 2.1. *conics of equation $ax^2 + cy^2 + dx = 1$. The solid-line curve is the hyperbola $x^2 - y^2 = 1$ ($a = 1$, $c = -1$, and $d = 0$). The dashed-line curve is the ellipse (circle here) $x^2 + y^2 = 1$ ($a = 1$, $c = 1$, and $d = 0$). The dotted-line curve is the parabole $x - y^2 = 1$ ($a = 0$, $c = -1$, and $d = 1$).*

Characteristic form of first-order equations

Quasi-linear first-order partial differential equations equations are linear in the differential terms and can cast put in the form:

$$P(x,\ y,\ u)\partial_x u + Q(x,\ y,\ u)\partial_y u = R(x,\ y,\ u). \tag{2.2}$$

The implicit solution can be written as $\psi(x,\ y,\ u(x,\ y)) = c$ (with c a constant). ψ is a first integral of the vector field $(P,\ Q,\ R)$. We have:

$$\partial_x \psi(x,\ y,\ u(x,\ y)) = 0 = \psi_x + \psi_u u_x,$$
$$\partial_y \psi(x,\ y,\ u(x,\ y)) = 0 = \psi_y + \psi_u u_y.$$

We can also write: $u_x = -\psi_x/\psi_u$ et $u_y = -\psi_y/\psi_u$. We can get a more symmetric relation:

$$P\psi_x + Q\psi_y + R\psi_u = 0,$$

which can be cast into a vector form, which is easier to interpret:

$$(P,\ Q,\ R) \cdot \nabla\psi = 0. \tag{2.3}$$

This means that at point M, the normal vector to the integral curve should be normal to the vector field $(P,\ Q,\ R)$. If the point O: $(x,\ y,\ u)$ and the neighboring point O': $(x + \mathrm{d}x,\ y + \mathrm{d}y,\ u + \mathrm{d}u)$ belong to the integral curve, then the vector $\mathbf{OO'}$: $(\mathrm{d}x,\ \mathrm{d}y,\ \mathrm{d}u)$ must be normal to $(P,\ Q,\ R)$: $\psi_x dx + \psi_y dy + \psi_u du = 0$. Since this must be true for any increment $\mathrm{d}x$, $\mathrm{d}y$, and $\mathrm{d}u$, we obtain the *characteristic equations*

$$\boxed{\frac{\mathrm{d}x}{P(x,\ y,\ u)} = \frac{\mathrm{d}y}{Q(x,\ y,\ u)} = \frac{\mathrm{d}u}{R(x,\ y,\ u)}} \tag{2.4}$$

Each pair of equations defines a curve in the space $(x,\ y,\ u)$. These curves define a two-parameter family (there are 3 equations, so 3 invariants but only 2 are independent): for

example, if p is a *first integral* of the first pair of equations, an integral surface of the first pair is given by an equation of the form $p(x,\ y,\ u) = a$, with a a constant. Similarly for the second pair: $q(x,\ y,\ u) = b$. The functional relation $F(a,\ b) = 0$ defines the integral curve.

Note that all solutions do not necessarily take the form $F(a,\ b) = 0$. This case is encountered, in particular, with *singular* solutions of differential equations.

Using characteristic equations can often help solve quasi-linear first order equations.

♣ **Example**. – We would like to find a solution to:

$$x\frac{\partial u}{\partial x} - y\frac{\partial u}{\partial y} = u^2.$$

Identifying P, Q, and R, we find: $P = x$, $Q = -y$, et $R = u^2$. The characteristic equation is

$$\frac{\mathrm{d}x}{x} = -\frac{\mathrm{d}y}{y} = \frac{\mathrm{d}u}{u^2}.$$

A first integral of the first pair is

$$\frac{\mathrm{d}x}{x} = -\frac{\mathrm{d}y}{y} \Rightarrow \ln x = -\ln y + \ln a,$$

with a a constant of integration. We then have $a = xy$. A first integral of the second pair is

$$-\frac{\mathrm{d}y}{y} = \frac{\mathrm{d}u}{u^2} \Rightarrow \ln y = \frac{1}{u} + b,$$

with b another constante of integration. We then get $b = \ln y - 1/u$. General solutions are of the form

$$F(a,\ b) = 0 \Rightarrow F\left(xy, \ln y - \frac{1}{u}\right) = 0.$$

This is the implicit solution to the equation. An explicit form is obtained by assuming that there is a function G such that $\ln y - 1/u = G(xy)$, that is,

$$u = \frac{1}{\ln y - G(xy)}.$$

Function G must be determined from the boundary conditions. □

Boundary conditions

In mechanics, we solve equations with space and time variables. In general, in order to work out a particular solution u to a partial differential equation, we need

- boundary conditions that specify how u varies along the domain boundaries at any time;
- the initial conditions that specify how varies u at the initial instant for any point in the domain.

We must solve what is called a boundary-value problem with initial conditions or, said differently, *initial boundary-value problem.* In some cases, we do not need as much information. For

example, for certain hyperbolic equations, one needs only the initial conditions while the elliptic problems require only boundary conditions (they generally reflect stationary processes).

We distinguish:

- *Dirichlet boundary conditions*: the boundary condition specifies the value u_0 that the function takes at a point or a curve
$$u(\mathbf{x}\ ;\ t) = u_0(t)$$
along a curve Γ.
- *Neuman boundary conditions*: the boundary conditions specify the derivative that the function takes at a point or a series of points. Physically, this reflects a flux condition through the domain boundary:
$$\frac{\partial u}{\partial \mathbf{n}}(\mathbf{n}\ ;\ t) = \boldsymbol{\phi}(t)$$
along a curve Γ, with $\mathbf{n}$ the normale to Γ and $\boldsymbol{\phi}(t)$ a flux function.

2.1.4 Variational equation

There is a mechanical principle known as the *variational principle* stating that if a process $J[u]$ (with J a functional and u a function) is steady and stable, then it should remain insensitive to small variations of u. This is written $\Delta J = 0$. A functional is a generalized function that involves both u and its derivatives (or integrals). For one-dimensional problems, a generic form of J is for example the form

$$J[u] = \int L(t,\ u,\ \dot{u}, \cdots)\mathrm{d}t, \tag{2.5}$$

with L a function of $u(t)$, t, and its derivatives.

For example, the Hamilton principle states that a particle moves so that the action integral which represents the difference between kinetic and potential energies is minimized

$$J = \int_{t_1}^{t_2} (\text{kinetic energy} - \text{potential energy})\mathrm{d}t.$$

A variational problem of the form $\delta J = 0$ with J given by equation (2.5) can be reduced to a purely differential equation. One can indeed show that $u(t)$ is solution to the following differential equation called the *Euler-Lagrange equation*

$$\frac{\partial L}{\partial u} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{u}}\right) + \frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\frac{\partial L}{\partial \ddot{u}}\right) + \cdots = 0.$$

♣ **Example**. – For instance, if y denotes the position of a point mass m linked to a spring of strength k, then $y(t)$ is determined by solving $\delta J = 0$, with

$$J = \frac{1}{2}\int_{t_1}^{t_2} (m\dot{y}^2 - ky^2)\mathrm{d}t.$$

After identifying the terms, we find $L(y,\ \dot{y}) = (m\dot{y}^2 - ky^2)/2$. We then deduce $L_y = -ky$ and $L_{\dot{y}} = m\dot{y}$. The resulting Euler-Lagrange equation is

$$-ky - \frac{\mathrm{d}}{\mathrm{d}t}m\dot{y} = 0 \Rightarrow \ddot{y} = -\frac{k}{m}y,$$

which is nothing but the Newton equation for an oscillatory mass. □

2.2 Equations in mechanics

We are going to see the main types of partial differential equations encountered in mechanics.

2.2.1 Convection equation

Convection is a mode of transfer of an element or a quantity that is advected by the fluid. For example, if a pollutant is released into a river, it is transported at the same speed as water. This is what we call *convection* or *advection* (convection is most often employed to describe thermal heat transfer).

The simplest convection equation is the following one

$$\frac{\partial f}{\partial t} + u\frac{\partial f}{\partial x} = 0, \tag{2.6}$$

where $f(x, t)$ is a quantity advected by a stream at constant velocity u. This is a first-order linear partial differential equation. The characteristic equation associated to the partial differential equation (2.6) is

$$\frac{\mathrm{d}x}{\mathrm{d}t} = u \text{ ou bien encore } \frac{\mathrm{d}x}{u} = \frac{\mathrm{d}t}{1} = \frac{\mathrm{d}f}{0}.$$

As u is assumed to be constant, this means that the solution of the characteristic equation is $x - ut = const$ any function $F(x - ut)$ whose argument is $x - ut$ is a solution of equation (2.6). A feature of this solution is that the original form $F(x)$ (at $t = 0$) is conserved in the course of movement: it is simply translated by ut as shown in Fig. 2.2.

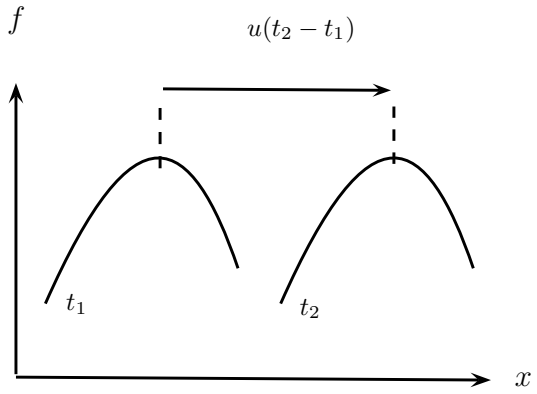


Figure 2.2. *advection of f.*

2.2.2 Diffusion equation (heat equation)

Diffusion is matter transport under the effect of thermal agitation (Brownian motion) or turbulence. In a stream, in addition to mean motion, there are also velocity fluctuations which disperse matter within the fluid volume.

One example is the classic diffusion equation of heat diffusion. The temperature $T(x,\ y\ ,\ t)$ varies over time within a material (dimension 2) according to the equation

$$\frac{\partial T}{\partial t} = \alpha \Delta T = \alpha \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} \right), \tag{2.7}$$

with $\alpha = k/(\varrho C)$ thermal diffusivity, ϱ density, k thermal conductivity, C specific heat.

A similar process occurs with matter. The one-dimensional diffusion equation is

$$\frac{\partial f}{\partial t} = D \frac{\partial^2 f}{\partial x^2}, \tag{2.8}$$

with D diffusion coefficient and where $f(x,\ t)$ is a function such as the pollutant concentration in a river. This is a second-order linear partial differential equation.

We talked so far about linear equations. In many problems of practical interest, the diffusion coefficient is not constant but depends on the function f. The diffusion equation is then nonlinear. For example, when a $D(f) = \kappa f^k$, the diffusion equation is

$$\frac{\partial f}{\partial t} = \kappa \frac{\partial}{\partial x} \left(f^k \frac{\partial f}{\partial x} \right), \tag{2.9}$$

with κ a constant and k an integer. For gas diffusion in a porous medium and groundwater flow, we have $k = 1$ (f represents the concentration); for the spreading of Newtonian fluid on a horizontal substrate, we have $k = 3$ (f is the height of fluid); for heat diffusion during the early stages of a nuclear explosion, we have $k = 5$.

Self-similar solution to the Green problem

For some initial conditions, Equation (2.8) admits analytical solutions in the form of self-similar solution $t^m F(\xi)$ with $\xi = x/t^n$. When we substitute f by this form into (2.8), we find that $n = \frac{1}{2}$. We note that m is not determined by the differential equation, but it is given the boundary conditions. Generally, in physical problems, we impose that the quantity of material released is constant

$$\int_{-\infty}^{\infty} f(x) \mathrm{d}x = V,$$

with V total volume (assumed to be constant with time). The change of variable gives $\int f(x) \mathrm{d}x = \int t^{m+1/2} F(\xi) \mathrm{d}\xi = V$. We then pose $m = -\frac{1}{2}$ since V does not depend on t.

The advantage of this change of variable is that it transforms the partial differential equation into a second-order ordinary linear differential equation, which is much easier to solve. Let us take a closer look at this transformation by considering a practical case: in a lake (water at rest), we release a volume V of pollutant, originally contained at point $x = 0$; the initial condition is therefore $f(x,\ 0) = \delta(x)$ where δ is the Dirac function ($\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ if $x \neq 0$). This problem where the initial condition is an impulse, i.e. a quantity localized at a point, is called *Green's problem*. Substituting the form $f = t^{-1/2} F(\xi)$ into Eq. (2.8), we obtain an ordinary differential equation for F and by doing so, we transform a partial differential problem in an ordinary differential equation:

$$F + \xi F'(\xi) + 2DF''(\xi) = 0,$$

which gives after integration

$$\xi F + 2DF' = a,$$

with a a constant of integration. Since the solution should be symmetric about $x = 0$ (thus $\xi = 0$), we have $F' = 0$ at $x = 0$ (F has a horizontal tangent at this point), hence $a = 0$. A new integration yields

$$F(\xi) = be^{-\frac{\xi^2}{4D}} \Rightarrow f(x,\, t) = \frac{b}{\sqrt{t}} e^{-\frac{x^2}{4Dt}},$$

with b another constant of integration. since $\int_{-\infty}^{\infty} e^{-\frac{x^2}{4D}} \mathrm{d}x = 2\sqrt{D\pi}$, we deduce $b = V/2\sqrt{D\pi}$. The solution reads

$$f(x,\, t) = \frac{V}{\sqrt{4\pi Dt}} e^{-\frac{x^2}{4Dt}}. \tag{2.10}$$

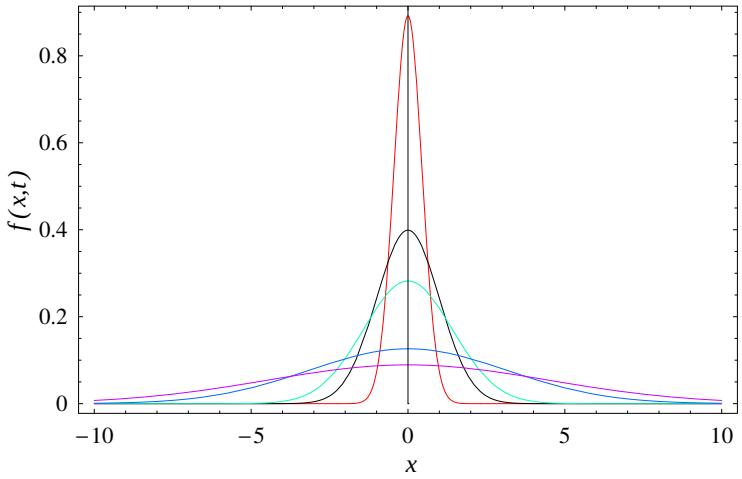


Figure 2.3. *diffusion of f. Calculation done with $D = 1$ m^2/s and at times $t = 0,1$, $t = 0,5$, $t = 1$, $t = 5$, and $t = 10$ s.*

As shown in Fig. 2.3, the shape of the diffusion front does not change with time (it is bell-shaped), although the front spreads. Note that the resulting solution is of great interest because it is the particular solution of the Green problem. For example, assume that the initial condition is more complex $f(x,\, 0) = g(x)$. Since the differential equation is linear, the sum of two solutions is also a solution. The general solution is then

$$f(x,\, t) = \frac{1}{\sqrt{4\pi Dt}} \int_{-\infty}^{\infty} g(\zeta) e^{-\frac{(x-\zeta)^2}{4Dt}} \mathrm{d}\zeta.$$

This means that the concentration f at any time t and for any x is the sum of elementary contributions induced by the distribution of source intensity $g(\zeta)$ per unit length.

Laplace transform

We wish to solve the diffusion equation:

$$\frac{\partial f}{\partial t} = D \frac{\partial^2 f}{\partial x^2}, \tag{2.11}$$

with D the diffusion coefficient (which is constant), subject to the following initial and boundary conditions

$$f(x,\ 0) = 0, \tag{2.12}$$
$$f(0,\ t) = a \text{ for } t > 0, \tag{2.13}$$
$$f(x,\ t) = 0 \text{ for } x \to \infty \text{ and } t > 0, \tag{2.14}$$

with a a constant. We transform it using the Laplace transform in t

$$\hat{f}(x,\ s) = \int_0^\infty e^{-st} f(x,\ t)\mathrm{d}t.$$

To transform (2.11), we just have to multiply it by e^{-st}, then integrate it from 0 to ∞ with respect to t. We get

$$\int_0^\infty De^{-st}\frac{\partial^2 f}{\partial x^2}\mathrm{d}t = D\frac{\partial^2}{\partial x^2}\int_0^\infty e^{-st} f\mathrm{d}t, \tag{2.15}$$
$$\int_0^\infty e^{-st}\frac{\partial f}{\partial t}\mathrm{d}t = \left[fe^{-st}\right]_0^\infty + \int_0^\infty se^{-st} f\mathrm{d}t, \tag{2.16}$$

where the term between brackets vanishes (given the initial condition). The Laplace transform of the linear diffusion equation (2.11) is

$$s\hat{f} = D\frac{\partial^2 \hat{f}}{\partial x^2}, \tag{2.17}$$

which, despite the partial derivatives, behaves like an ordinary differential equation in x. The Laplace transform of the boundary conditions (2.13) and (2.14) yields

$$\hat{f}(0,\ s) = \int_0^\infty ae^{-st} = \frac{a}{s}, \tag{2.18}$$
$$\hat{f}(x,\ s) = 0 \text{ when } x \to \infty. \tag{2.19}$$

The solution to (2.17) is then

$$\hat{f}(x,\ s) = \frac{a}{s}\exp\left(-x\sqrt{\frac{s}{D}}\right),$$

whose inverse Laplace transform is

$$f(x,\ t) = a\left(1 - \mathrm{Erf}\left(\frac{x}{2\sqrt{Dt}}\right)\right),$$

with Erf the error function. Figure 2.4 shows f profiles at different times. At long time, we get

$$\lim_{t\to\infty} f(x,\ t) = a,$$

whose profile tends to a uniform profile $f = a$.

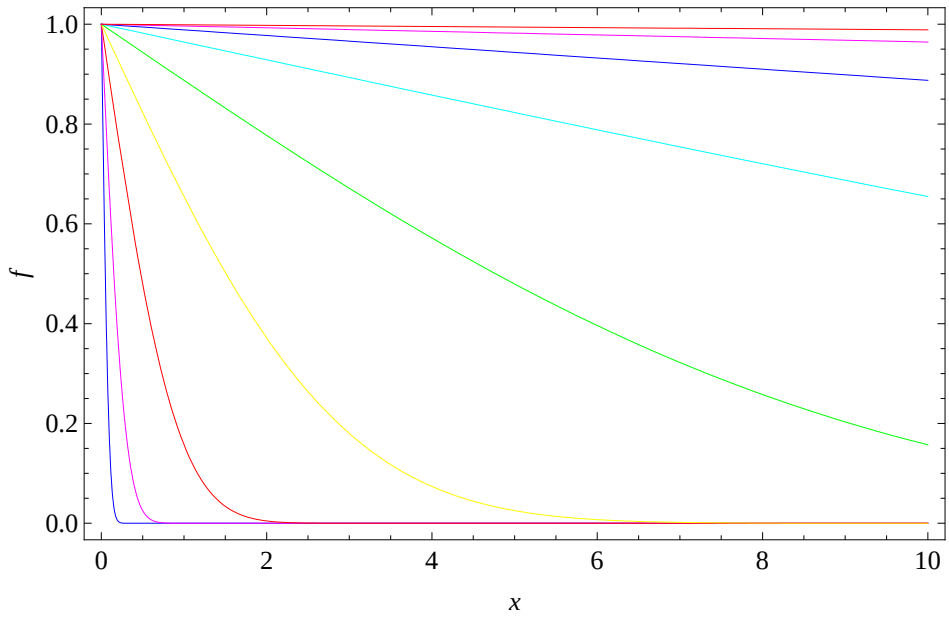


Figure 2.4. *variation in f for $t = 10^{-2}$, 10^{-2}, 1, $10^{+1}\ldots$, 10^{+6}. Calculation done for $a = 1$ and $D = 1$ m^2 /s.*

2.2.3 Convection-diffusion equation

Convection-diffusion is a combination of two phenomena. This is the phenomenon commonly encountered in hydraulics. For example, the discharge of a pollutant into a river leads to a transport of this pollutant by diffusion (due to turbulence) and convection (advection speed of the water). The characteristic equation is

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\partial f}{\partial t} + u\frac{\partial f}{\partial x} = D\frac{\partial^2 f}{\partial x^2}, \tag{2.20}$$

where D and u are assumed constant. This equation can be reduced to a linear diffusion problem by making the following change of variable (which amounts to changing the reference frame and placing us in the frame attached to the stream)

$$\begin{aligned} \zeta &= x - ut, \\ \tau &= t. \end{aligned}$$

We get

$$\begin{aligned} \frac{\partial \cdot}{\partial x} &= \frac{\partial \cdot}{\partial \zeta}\frac{\partial \zeta}{\partial x} + \frac{\partial \cdot}{\partial \tau}\frac{\partial \tau}{\partial x}, \\ &= \frac{\partial \cdot}{\partial \zeta}, \\ \frac{\partial \cdot}{\partial t} &= \frac{\partial \cdot}{\partial \zeta}\frac{\partial \zeta}{\partial t} + \frac{\partial \cdot}{\partial \tau}\frac{\partial \tau}{\partial t}, \\ &= -u\frac{\partial \cdot}{\partial \zeta} + \frac{\partial \cdot}{\partial \tau}. \end{aligned}$$

Equation (2.20) becomes

$$\frac{\partial f}{\partial \tau} = D\frac{\partial^2 f}{\partial \zeta^2},$$

which is similar to the linear diffusion equation (2.8) seen above.

A special case of convection-diffusion is met with the Burgers equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = D\frac{\partial^2 u}{\partial x^2}, \tag{2.21}$$

which can also be transformed into a diffusion equation using the Cole-Hopf transformation:

$$u = -\frac{2D}{\phi}\frac{\partial \phi}{\partial x},$$

with $\phi(x,\, t)$ an auxiliary fonction. Indeed we have

$$\begin{aligned}
\frac{\partial u}{\partial x} &= -\frac{2D}{\phi}\frac{\partial^2 \phi}{\partial x^2} + \frac{2D}{\phi^2}\left(\frac{\partial \phi}{\partial x}\right)^2, \\
\frac{\partial u}{\partial t} &= -\frac{2D}{\phi}\frac{\partial^2 \phi}{\partial x \partial t} + \frac{2D}{\phi^2}\frac{\partial \phi}{\partial x}\frac{\partial \phi}{\partial t}, \\
\frac{\partial^2 u}{\partial x^2} &= -\frac{2D}{\phi}\frac{\partial^3 \phi}{\partial x^3} - \frac{4D}{\phi^3}\left(\frac{\partial \phi}{\partial x}\right)^3 + \frac{6D}{\phi^2}\frac{\partial^2 \phi}{\partial x^2}\frac{\partial \phi}{\partial x}.
\end{aligned}$$

After simplification, we obtain

$$\frac{\partial \phi}{\partial t}\frac{\partial \phi}{\partial x} - \phi\frac{\partial^2 \phi}{\partial x \partial t} + 2D\left(\phi\frac{\partial^3 \phi}{\partial x^3} - \frac{\partial \phi}{\partial x}\frac{\partial^2 \phi}{\partial x^2}\right) = 0,$$

which can be transformed—by dividing it by ϕ^2, then integrating with respect to x, and ultimately by multiplying it again by ϕ—into a linear diffusion equation

$$\frac{\partial \phi}{\partial t} = D\frac{\partial^2 \phi}{\partial x^2}.$$

2.2.4 Wave

Dynamic waves are solutions to a differential equation such as the following (second-order) partial differential equation:

$$\frac{\partial^2 \phi}{\partial t^2} = c^2\frac{\partial^2 \phi}{\partial x^2}, \tag{2.22}$$

with c the (phase) velocity. This form is not exhaustive; for example, the equation of gravity waves reads

$$\frac{\partial^2 \phi}{\partial t^2} = -g\frac{\partial \phi}{\partial y},$$

with here ϕ the velocity potentiel ($\mathbf{u}(x,\, y,\, t) = \nabla\phi$) and g gravity acceleration.

Often solutions are sought in the form of harmonics (periodic wave)

$$\phi(t) = A\exp[\imath(kx - \omega t)] = \mathrm{Re}(A)\cos(kx - \omega t) - \mathrm{Im}(A)\sin(kx - \omega t),$$

where A is the *amplitude*, k *wave number* ($\lambda = 2\pi/k$ *wavelength*), ω *angular frequency*; we also introduce a frequency f defined as $f = \omega/(2\pi)$: the number of complete oscillations during a second at a given position. The period is defined as $T = \lambda/c$.

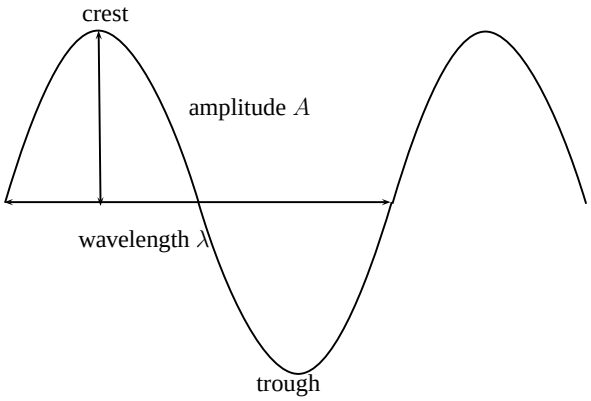


Figure 2.5. *wave length and amplitude of harmonic wave.*

The *wave velocity* is $c = \omega/k$. This means that for an interval δt, the wave has moved a distance $c\delta t$. The dispersion relation $\omega(k)$ is here linear because we have: $\omega(k) = ck$, i.e. the wave crests move at a constant speed, which is independent of the wavelength. In most systems, the relation is not linear, which in practice means that the crest velocity depends on the wavelength. We then introduce the phase velocity c_p

$$c_p = \frac{\omega(k)}{k}.$$

In a physical process where waves result from the superposition of many harmonic waves of different wavelength, each harmonic component moves at its own speed, which ultimately leads to a separation or *dispersion* of the wave, hence the name *dispersion relation* for $\omega(k)$. There is a third velocity, called *group velocity*, which represents the speed at which the energy associated with the wave propagates:

$$c_g = \frac{\mathrm{d}\omega}{\mathrm{d}k}. \tag{2.23}$$

In general for most physical processes, we have $c_g \leq c_p$.

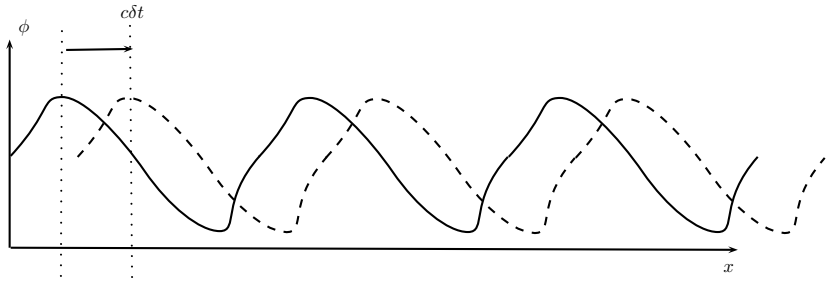


Figure 2.6. *rightward shift of a forward wave a speed c .*

The differential equation (2.22) is linear, which implies that any combination of solutions is also solution (superposition principle). There are two directions of propagation:

- forward wave $f = f(x - ct)$: the wave goes in the $x > 0$ direction;
- backward wave $f = f(x + ct)$: the wave goes in the $x < 0$ direction.

Equation (2.22) can also be cast in the form of products

$$\frac{\partial^2 f}{\partial t^2} - c^2 \frac{\partial^2 f}{\partial x^2} = \left(\frac{\partial}{\partial t} - c \frac{\partial}{\partial x} \right) \left(\frac{\partial}{\partial t} + c \frac{\partial}{\partial x} \right) f = 0,$$

which makes it possible to transform a second-order partial differential equation into a system of first-order equations:

$$\begin{cases} f_t - c f_x = v, \\ v_t + c v_x = 0. \end{cases}$$

The general solution to the wave equation (2.22) reads

$$f = a(x - ct) + b(x + ct),$$

with a and b two functions. This is the *d'Alembert* solution.

Note that in many cases of practical interest, the equations are linear, the linearity allows to apply the principle of superposition. A stationary wave results from the superimposition of a backward wave and a forward wave of same amplitude. In this case, time dependence is removed.

2.2.5 Laplace equation

Elliptic equations generally reflect how a process at equilibrium is spatially organized. The prototype of the elliptic equation is the *Laplace* equation:

$$u_{xx} + u_{yy} = 0. \tag{2.24}$$

For example, the heat equation (2.7) in steady state ($\partial_t T = 0$) becomes elliptical. The Laplace equation is used to describe a large number of steady flows in environmental problems. For instance, the slow flow of water in a porous medium is also a Laplace equation. Indeed, if the velocity $\mathbf{u}$ follows Darcy's law, then it is connected to the pressure gradient p by: $\mathbf{u} = -k\nabla p/\mu$ with μ the viscosity and k the permeability of the medium. We can recast this equation as follows $\mathbf{u} = -\nabla\psi$ with $\psi = -kp/\mu$. We say that $\mathbf{u}$ is derived from the potential ψ. The continuity equation (incompressible fluid) requires that $\operatorname{div} \mathbf{u} = 0$, hence

$$\nabla \cdot \nabla\psi = 0 \Rightarrow \Delta\psi = 0.$$

Exercises

Exercise 2.1 Calculate the phase velocity and group velocity of the following equation:

$$u_t + u_x + u_{xxx} = 0.$$

$\rightsquigarrow$ *Réponse* :

We seek harmonic solutions of the form:

$$u(x,\ t) = A\mathrm{e}^{\imath(kx-\omega t)}.$$

The dispersion relation is:

$$\omega = k - k^3.$$

We deduce that the phase velocity is

$$c = \frac{\omega}{k} = 1 - k^2,$$

i.e. a function of k. This is thus a dispersive wave. The group velocity

$$c_g = \frac{\mathrm{d}\omega}{\mathrm{d}k} = 1 - 3k^2,$$

which is lower than the phase velocity, as expected. □

Exercise 2.2 Consider the following initial-value problem

$$t\frac{\partial u}{\partial x} - x\frac{\partial u}{\partial t} = 0,$$

with $u(x,\ 0) = f(x)$ for $x > 0$. What type is this equation? Solve it after determining the associate characteristic equation.

Exercise 2.3 A function $u(x,\ t)$ decreases like

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -k,$$

along a curve $x = x_s(t)$ in the $x - t$ plane. Which partial differential equation does u satisfy? Conversely, show that any partial differential equation of the form

$$\frac{\partial u}{\partial t} + a(u,\ x,\ t)\frac{\partial u}{\partial x} = b(u,\ x,\ t),$$

can be transformed into a first-order ordinary differential equation

$$\frac{\mathrm{d}u}{\mathrm{d}t} = g$$

along a curve whose equation has to be specified.

Exercise 2.4 The Euler-Darboux equation reads

$$u_{xy} + \frac{au_x - bu_y}{x - y} = 0.$$

Characterize this equation.

Exercise 2.5 The Helmholtz equation reads

$$\nabla^2 u + k^2 u = 0.$$

Characterize this equation.

Exercise 2.6 The Klein-Gordon equation is a variant of the Schrödinger equation, which described how an elementary particle behaves. It reads

$$\frac{\partial^2 u}{\partial t^2} - \gamma^2 \frac{\partial^2 u}{\partial x^2} + c^2 u = 0.$$

Characterize this equation. Seek periodic solutions in the form $u(x,\ t) = a(k)\exp(\imath kx + \lambda(k)t)$ with a the amplitude of the wave and where λ and k are the modes. Determine the mode λ? Is the solution stable?

Exercise 2.7 Consider the equation:

$$\dot{y} = -\frac{y(y^2 - x)}{x^2},$$

with initial condition $y(1) = 0$. Answer the following equations:

- which type is this equation?
- solve it numerically;
- put the equation in the form $A\mathrm{d}x + B\mathrm{d}y = 0$. Is it an exact differential?
- multiply the equation above $2/(2xy^3 - x^2 y)$. Is it an exact differential?
- integrate the equation and compare with the numerical solution.

Exercise 2.8 Find the function $y(x)$ that minimize the functional

$$J[y] = \int_0^1 (y^2 + y'^2)\mathrm{d}x,$$

hiven that $y(0) = 0$ and $y(1) = 1$.

3 Analytical tools

3.1 Overview

Several methods are available for solving differential equations. If there is no universal solving technique that can solve any type of equation, there are a number of methods that work in most cases of practical interest. Exact methods include:

- *variable separation*: this technique can transform a partial differential equation into a set of differential equations;
- *integral transformation*: Fourier or Laplace transform can be used to transform a PDE into a linear ordinary differential equation when the domain of integration is infinite (or semi-infinite). An example is given with the diffusion equation in § 2.2.2;
- *Green method*: for linear equations with boundary conditions that are also linear, it is possible to exploit the linearity by first seeking to solve a Green problem, that is, the same differential equation but with boundary conditions involving impulse (Dirac) functions. The final solution is obtained by summing the response to each elementary impulse. An example is given with the diffusion equation in § 2.2.2;
- *invariance group*: the idea is to exploit geometric transformations (forming what is called a group in mathematics) which leave an equation invariant. Among the most frequent, the translation invariance and stretching group can find self-similar solutions. These methods simplify the problem by transforming the partial differential equation into an ordinary differential equation;
- *hodograph method*: some equations are simpler to solve when the role of dependent and independent variables is swapped;
- *eigenfunction expansion*: the solution to a linear differential equation (with linear boundary conditions) is sought in the form of an infinite series of eigenfunctions.

For some equations, there are specific methods that we do not describe here. For example, conformal transformations offer an application of the theory of complex-valued functions to solve the Laplace equation.

Approximate methods include:

- *pertubation techniques*: we transform a nonlinear problem into a (hierarchical) set of linear equations that can approximate the nonlinear equation;

- *asymptotic methods*: we simplify the equations by removing terms whose order of magnitude is small compared to other terms;
- *numerical methods*: equations are discretized and solved by iterative methods using a computer. Other numerical methods: Galerkin methods seek numerical solutions by decomposing the form of known functions (spline, polynomial, wavelet, etc.).

3.1.1 Perturbation techniques

It is quite common in mechanics to obtain rather complex differential equations, whose some terms are weighted by coefficients that take values relatively low compared to other contributions. The idea is then

- to approximate the solution by a series of functions, whose order of magnitude decreases;
- to substitute this expression into the original equation;
- to group terms of same order to form a hierarchy of equations;
- to solve the equations iteratively.

♣ **Example**. – Let us take an example with this second-order differential equation

$$y'' + \epsilon y' + y = 0, \tag{3.1}$$

with the initial conditions $y(0) = 1$ and $y'(0) = 0$; ϵ is much smaller than unity (e.g., $\epsilon = 0.1$). We consider the following expansion in ϵ

$$y(x) = y_0(x) + \epsilon y_1(x) + \epsilon^2 y_2 + \ldots \epsilon^n y_n + \ldots,$$

with y_k a function of x such that $O(y_k) = 1$ over the interval $[0,\ 1]$. This expression is substituted into equation (3.1), which gives

$$\begin{aligned}(y_0(x) + \epsilon y_1(x) + \epsilon^2 y_2 + \ldots \epsilon^n y_n + \ldots)''+\\ \epsilon(y_0(x) + \epsilon y_1(x) + \epsilon^2 y_2 + \ldots \epsilon^n y_n + \ldots)'+\\ (y_0(x) + \epsilon y_1(x) + \epsilon^2 y_2 + \ldots \epsilon^n y_n + \ldots) = 0\end{aligned}$$

The boundary conditions yield

$$\begin{aligned}y_0(0) + \epsilon y_1(0) + \epsilon^2 y_2(0) + \ldots \epsilon^n y_n(0) + \ldots = 1,\\ y_0'(0) + \epsilon y_1'(0) + \epsilon^2 y_2'(0) + \ldots \epsilon^n y_n'(0) + \ldots = 0.\end{aligned}$$

To order ϵ^0, we obtain

$$y_0'' + y_0 = 0,$$

with boundary conditions $y_0(0) = 1$ and $y_0'(0) = 0$. Integrating this equation gives: $y_0(x) = \cos x$. To ϵ^1, we get

$$y_1'' + y_1 = -y_0',$$

with boundary conditions $y_1(0) = 0$ and $y_1'(0) = 0$ and whose integration gives: $y_1(x) = \frac{1}{2}(\sin x - x\cos x)$. This computation can be iterated indefinitely. The solution to order $O(\epsilon^2)$ is

$$y = \cos x + \frac{1}{2}\epsilon(\sin x - x\cos x) + O(\epsilon^2).$$

Figure 3.1 shows good agreement between exact and approximate solutions. □

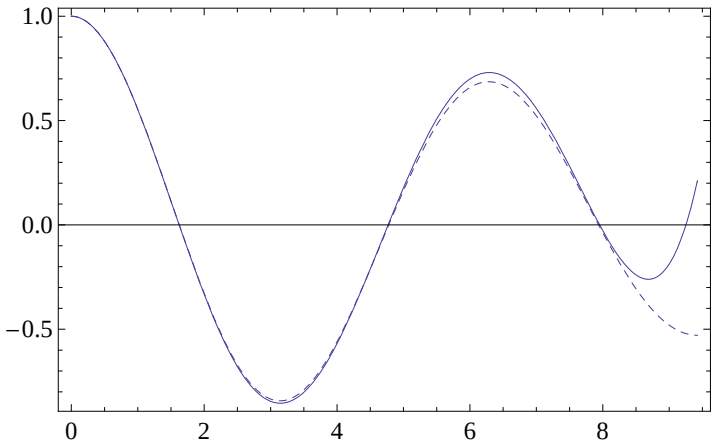


Figure 3.1. *comparison between the exact solution (solid line) and approximate solution to order 2 (dashed line) of equation (3.1) with* $\epsilon = 0.1$.

3.1.2 Asymptotic methods

In equations where several contributions come up, it is rare that all terms have the same importance. In seeking what are the dominant terms, we may obtain an asymptotic solution towards which the true solution tends. In general, we seek a balance between two, exceptionally three terms.

♣ **Example**. – Let us consider the differential equation

$$y'' + xy' + y = 0, \tag{3.2}$$

with initial conditions: $y(0) = 1$ and $y'(0) = 0$. The solution is $y = \exp(-x^2/2)$. We would like to approximate the solution for $x \to 0$ without using our knowledge of the true solution. To do this we will examine the contributions of the equation two by two:

- let us assume that $y'' \ll y$. We have to solve $x\tilde{y}' + \tilde{y} = 0$, whose first integral is $x\tilde{y} = a$, with a a constant. It is not possible to satisfy the boundary conditions. This balance is not possible;

- let us assume that $y \ll y''$. The dominant balance is $\tilde{y}'' + x\tilde{y}' = 0$, whose solution is $\tilde{y} = 1$. The assumption $y \ll y''$ is not satisfied, which means that this balance does not make sense;

- the only possibility is then $xy' \ll y$, which leads us to $\tilde{y}'' + \tilde{y} = 0$, whose solution is $\tilde{y} = \cos x$. We deduce that $x\tilde{y}' = -x \sin x$ is smaller than $\tilde{y}$ when $x \to 0$.

The approximate solution to equation (3.2) is $\tilde{y} = \cos x$, which provides a relatively accurate representation of the solution when $x \to 0$, as shown in Figure 3.2. □

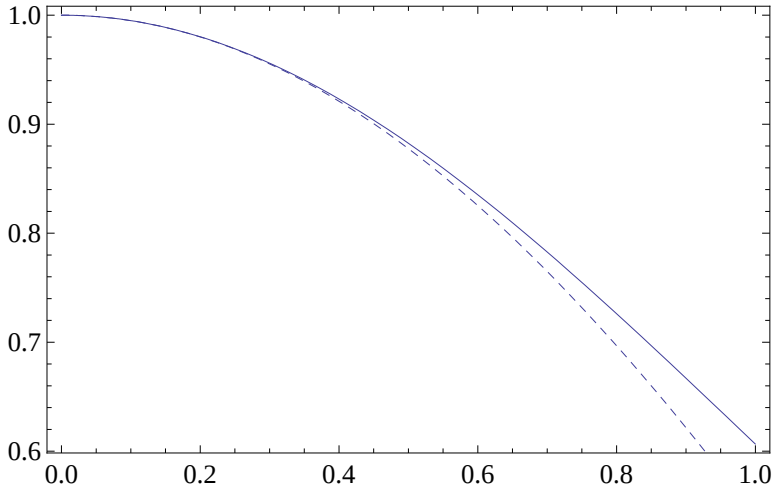


Figure 3.2. *Comparison between the exact solution* $y = \exp(-x^2/2)$ *(solid line) and approximate* $\tilde{y} = \cos x$ *(dashed line) of Equation (3.2).*

3.1.3 Similarity solutions

Here we will see two techniques for working out self-similar solutions to a partial differential equation (if such solutions exist):

- in the first method, we will see that when dimensional analysis of the partial differential equation and of its boundary/initial conditions shows that there are only two dimensionless numbers that define the problem, i.e. when the solution can be written as $\Pi_1 = \phi(\Pi_2)$, then we can work out a self-similar solution;
- in the second method, we first make equations dimensionless, then ask whether they are invariant to a "stretching" transformation. In this case, we can reduce the order of the partial differential equation and transform partial differential equations into ordinary equations, which are far easier to solve.

These two methods are studied through the example of the heat equation (see also § 2.2.2).

Insight from dimensional analysis

Let us consider the heat equation (2.7) in dimension 1 for a bar of section S

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}, \tag{3.3}$$

with α thermal diffusion, $T(x,\, t)$ temperature, x abscissa in the bar direction. Thermal energy E is conserved

$$\int_{-\infty}^{\infty} T(x,\ t)\mathrm{d}x = V = \frac{E}{cS}, \tag{3.4}$$

with c heat capacity. There are $n = 5$ variables: T, x, t, α, and V; the other variables (E, c, and S) are introduced through V.

The dimensional matrix is the following

	T	x	t	α	V
homogeneous to	K	m	s	m^2 /s	m· K
power decomposition :					
power of m	0	1	0	2	1
power of s	0	0	1	-1	0
power of K	1	0	0	0	1

This is a 3×5 matrix of rank 3 (the fourth column is obtained by linear combination of columns 2 and 3, column 5 is the sum of columns 1 and 2). We can therefore form $k = n - r = 2$ dimensionless numbers. Let

$$\Pi_1 = x\alpha^a t^b V^c \text{ et } \Pi_2 = T\alpha^{a'} t^{b'} V^{c'}.$$

To get $[\Pi_1] = 0$, we must have

$$[\text{m } (\text{m}^2/\text{s})^a \text{ s}^b \text{ (mK)}^c] = 0,$$

which leads to the following system of equations

$$\begin{aligned} \text{for m} &: 0 = 2a + c + 1, \\ \text{for s} &: 0 = -a + b, \\ \text{for K} &: 0 = c, \end{aligned}$$

whose solution is $a = -\frac{1}{2}$, $b = -\frac{1}{2}$, and $c = 0$. The first dimensionless group is

$$\Pi_1 = \frac{x}{\sqrt{\alpha t}}.$$

To get $[\Pi_2] = 0$, we must have

$$[\text{K } (\text{m}^2/\text{s})^{a'} \text{ s}^{b'} \text{ (mK)}^{c'}] = 0,$$

which leads to the following system of equations

$$\begin{aligned} \text{for m} &: 0 = 2a' + c', \\ \text{for s} &: 0 = -a' + b', \\ \text{for K} &: 0 = c' + 1, \end{aligned}$$

whose solution is $a' = \frac{1}{2}$, $b' = \frac{1}{2}$, and $c' = -1$. The second dimensionless group is

$$\Pi_2 = \frac{T\sqrt{\alpha t}}{V}.$$

Dimensional analysis leads us to pose the solution in the form $\Pi_2 = F(\Pi_1)$. We substitute T into (3.3), with T defined by

$$T = \frac{V}{\sqrt{\alpha t}} F(\xi),$$

with $\xi = x/\sqrt{\alpha t}$. We get

$$\begin{aligned}\frac{\partial T}{\partial t} &= -\frac{1}{2}\frac{V}{t^{3/2}\sqrt{\alpha}}F(\xi) - \frac{1}{2}\xi\frac{V}{t^{3/2}\sqrt{\alpha}}F'(\xi)\\ \frac{\partial T}{\partial x} &= \frac{V}{t\alpha}F'(\xi),\\ \frac{\partial^2 T}{\partial x^2} &= \frac{V}{(t\alpha)^{3/2}}F''(\xi),\end{aligned}$$

which leads to write the heat equation in the form of a second-order ordinary differential equation

$$-\frac{1}{2}F - \frac{1}{2}\xi F' = F'',$$

which is easy to integrate

$$\frac{1}{2}\xi F + F' = a_0,$$

with a_0 a constant of integration. If propagation occurs in both directions $x \to \infty$ et $x \to -\infty$, then the solution is even and for $\xi = 0$, $F' = 0$ (horizontal tangent). Eventually we get $a_0 = 0$. A new integration leads to

$$\frac{F'}{F} = -\frac{1}{2}\xi \Rightarrow F = a_1 \exp\left(-\frac{1}{4}\xi^2\right)$$

with a_1 a constant of integration. Using Equation (3.4) and since $\int_{\mathbb{R}} F \mathrm{d}\xi = 1$, we deduce $a_1 = 1/(2\sqrt{\pi})$.

The solution reads

$$T = \frac{V}{2\sqrt{\pi\alpha t}} \exp\left(-\frac{1}{4}\frac{x^2}{\alpha t}\right).$$

Mathematical derivation of similarity solution

We first make Equation (3.3) dimensionless by introducing dimensionless variables

$$\begin{aligned}T &= T_*\hat{T},\\ t &= \tau_*\hat{t},\\ x &= L_*\hat{x},\end{aligned}$$

with T_*, τ_*, and L_* temperature, time, and length scales; $\hat{T}$, $\hat{t}$, and $\hat{x}$ are dimensionless temperature, time, and length. When substituting these variables into (3.3), we get $L_*^2 = \alpha\tau_*$ while the boundary condition (3.4) imposes $L_*T_* = V$. A third condition is needed to determine all scales; we assume that τ_* is known, which makes it possible to derive the two other scales above.

When written in a dimensionless form, (3.3) and (3.4) read

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2}, \tag{3.5}$$

$$\int_{-\infty}^{\infty} T(x,\ t)\mathrm{d}x = 1, \tag{3.6}$$

where the hat has been removed for the sake of simplicity.

A solution to a partial differential equation of the form $G(x,\ t,\ T) = 0$ is said to to be *self similar* if we can find a set of coefficients a and b such that for any scalar λ we have $G(\lambda x,\ \lambda^a t,\ \lambda^b T) = 0$. This means that the solution $T(x,\ t)$ to the equation $G = 0$ is invariant when we stretch the variables by multiplying them by a factor of proportionality. Let us seek these factors by considering the following stretching, whose intensity depends on the parameter λ:

$$\begin{aligned} x &\to \lambda x', \\ t &\to \lambda^a t', \\ T &\to \lambda^b T', \end{aligned}$$

with a and b two constants to be determined. We substitute these expressions into the heat equation (3.5), which yields

$$\frac{\lambda^b}{\lambda^a}\frac{\partial T'}{\partial t'} = \frac{\lambda^b}{\lambda^2}\frac{\partial^2 T'}{\partial x'^2}. \tag{3.7}$$

This equation is identical to Equation (3.5) if we take $a = 2$. The boundary condition (3.6) gives

$$\int_{-\infty}^{\infty} \lambda^b \lambda T(x,\ t)\mathrm{d}x = 1, \tag{3.8}$$

which imposes $b = -1$.

It can be shown that the solutions invariant to the *stretching* transformation is then given by the characteristic equation

$$\frac{\mathrm{d}x}{x} = \frac{\mathrm{d}t}{at} = \frac{\mathrm{d}T}{bT}. \tag{3.9}$$

☣ **Proof**. If a solution is self-similar, then we have $G(\lambda x,\ \lambda^a t,\ \lambda^b T) = 0$. Let us differentiate this equation with respect to λ and set $\lambda = 1$; we deduce the relation:

$$x'\frac{\partial G}{\partial x'} + at'\frac{\partial G}{\partial t'} + bT'\frac{\partial G}{\partial T'} = 0.$$

The geometric interpretation is simple: the vector ∇G is perpendicular to the vector $(x',\ at',\ bT')$. If a point M of coordinates $(x',\ t',\ T')$ lies on the solution surface, then a nearby point M' $(x' + \mathrm{d}x',\ t' + \mathrm{d}t',\ T' + \mathrm{d}T')$ must also be on the surface and the increment vector between M and M' $(\mathrm{d}x',\ \mathrm{d}t',\ \mathrm{d}T')$ must also be normal to the solution surface, since to first order we have

$$G(x' + \mathrm{d}x',\ t' + \mathrm{d}t',\ T' + \mathrm{d}T') = 0,$$

A first-order series expansion gives

$$\mathrm{d}x'\frac{\partial G}{\partial x'} + \mathrm{d}t'\frac{\partial G}{\partial t'} + \mathrm{d}T'\frac{\partial G}{\partial T'} = 0.$$

Comparing both equations leads to deducing that $(\mathrm{d}x',\ \mathrm{d}t',\ \mathrm{d}T')$ and $(x',\ at',\ bT')$ are parallel. Equation (3.9) simply expresses the parallelism condition between both vectors. This may seem more complex than the original equation since it replaces a partial differential equation by a system of 3 differential equations. In fact, we have made our life easier because we know how to solve the above equations two by two. □

The characteristic equation associated with Equation (3.5) is

$$\frac{\mathrm{d}x}{x} = \frac{\mathrm{d}t}{2t} = -\frac{\mathrm{d}T}{T},$$

which admits two first integral: $\xi = x/t^{1/2}$ (obtained with the left-hand terms) and $\tau = Tt^{1/2}$. Similarity solutions are of the form $\tau(\xi)$:

$$T = \frac{1}{\sqrt{t}} H(\xi).$$

Substituting this expression into Equation (3.5), we find

$$-\frac{1}{2}H - \frac{1}{2}\xi H' = H'',$$

whose solution is $a_2 \exp\left(-\frac{1}{4}\xi^2\right)$, with a_2 a constant of integration. The boundary condition (3.6) provides the value: $a_2 = 1/(2\sqrt{\pi})$. The dimensionless solution is

$$\hat{T} = \frac{1}{2\sqrt{\pi \hat{t}}} \exp\left(-\frac{1}{4}\frac{\hat{x}^2}{\hat{t}}\right),$$

while its dimensional form is

$$T = \frac{V}{2\sqrt{\pi \alpha t}} \exp\left(-\frac{1}{4}\frac{x^2}{\alpha t}\right).$$

Summary

The first method is used to build self-similar solutions (when they exist) step by step and has the advantage of being a physical approach, but requires a lot of work. The second method, somewhat more mathematical, can quickly determine whether self-similar solutions exist and, when appropriate, to work it out.

In practice, if we consider a partial differential equation of the form $F(u,\ x,\ t)$ with u the dependent variable x and t independent variables, we use a *stretching* transformation for a parameter λ:

$$u \to u' = \lambda^{\alpha} u, \tag{3.10}$$

$$t \to t' = \lambda^{\beta} t, \tag{3.11}$$

$$x \to x' = \lambda x. \tag{3.12}$$

where α and β are two constants to be determined; they are determined by substituting these expressions into equation $F(u,\ x,\ t)$ and the boundary/initial conditions and then by seeking for what values of α and β, the transformed equations are independent of λ. Once these parameters have been found, we can solve the *characteristic equation*

$$\frac{\mathrm{d}x}{x} = \frac{\mathrm{d}t}{\beta t} = \frac{\mathrm{d}u}{\alpha u}.$$

This equation shows that the self-similar solution of $F(u,\ x,\ t)$ can be written as

$$u(x,\ t) = t^{\alpha/\beta} f(x/t^{1/\beta}). \tag{3.13}$$

3.2 First-order hyperbolic differential equations

In fluid mechanics, we are often faced with hyperbolic equations or systems of n hyperbolic equations:

- dimension 1: nonlinear convection equation, for example the kinematic wave equation, which describes flood propagation in rivers

$$\frac{\partial h}{\partial t} + K\sqrt{i}\frac{\partial h^{5/3}}{\partial x} = 0,$$

  with h flow depth, K Manning-Strickler coefficient, et i bed gradient;

- dimension 2: Saint-Venant equations

$$\frac{\partial h}{\partial t} + \frac{\partial h\bar{u}}{\partial x} = 0, \tag{3.14}$$
$$\frac{\partial \bar{u}}{\partial t} + \bar{u}\frac{\partial \bar{u}}{\partial x} = g\sin\theta - g\cos\theta\frac{\partial h}{\partial x} - \frac{\tau_p}{\varrho h}, \tag{3.15}$$

  with $\bar{u}$ flow-depth averaged velocity, h flow depth, θ bed slope, τ_p bottom shear stress;

- dimension 3 : Saint-Venant equations with advection of pollutant

$$\frac{\partial h}{\partial t} + \frac{\partial h\bar{u}}{\partial x} = 0, \tag{3.16}$$
$$\frac{\partial \bar{u}}{\partial t} + \bar{u}\frac{\partial \bar{u}}{\partial x} = g\sin\theta - g\cos\theta\frac{\partial h}{\partial x} - \frac{\tau_p}{\varrho h}, \tag{3.17}$$
$$\frac{\partial \varphi}{\partial t} + \bar{u}\frac{\partial \varphi}{\partial x} = 0, \tag{3.18}$$

  with φ pollutant concentration.

All these differential equations are first order and are evolution equations. Here we will focus on problems with one space variable x, but what we will say can be generalized to two (or more) space variables.

The key element in solving hyperbolic differential equations hinges upon the concept of information. Through the example of the wave equation and the convection equation, we have already seen that a partial differential equation expresses a physical process in which information spreads. The questions that arise are: in which direction does this information propagate? Is information conserved or attenuated? The answer these questions, we will go through the notions of *characteristic curve* (information propagation) and *Riemann variables* (quantity of information conveyed).

3.2.1 Characteristic equation

Let us first consider the following advection equation with $n = 1$ space variable:

$$\partial_t u(x,\ t) + a(u)\partial_x u(x,\ t) = 0, \tag{3.19}$$

subject to one boundary condition of the form:

$$u(x,\ 0) = u_0(x) \text{ at } t = 0, \tag{3.20}$$

A characteristic curve is a curve $x = x_c(t)$ along which the partial differential equation $\partial_f U + a\partial_x U = 0$ is equivalent to an ordinary differential equation. Consider a solution $u(x,\ t)$ of the differential system. Along the curve $\mathcal{C}$ of equation $x = x_c(t)$ we have: $u(x,\ t) = u(x_c(t),\ t)$ and the rate change is:

$$\frac{\mathrm{d}u(x_c(t),\ t)}{\mathrm{d}t} = \frac{\partial u(x,\ t)}{\partial t} + \frac{\mathrm{d}x_c}{\mathrm{d}t}\frac{\partial u(x,\ t)}{\partial x}.$$

Suppose now that the curve $\mathcal{C}$ satisfies the equation $\mathrm{d}x_c/\mathrm{d}t = a(u)$. So we immediately obtain:

$$\frac{\mathrm{d}u(x,\ t)}{\mathrm{d}t} = \frac{\partial u(x,\ t)}{\partial t} + a\frac{\partial u(x,\ t)}{\partial x} = 0. \tag{3.21}$$

Since $\mathrm{d}u(x,\ t)/\mathrm{d}t = 0$ along $x_c(t)$, this means that $u(x,\ t)$ is conserved along this curve. Since u is constant $a(u)$ is also constant, so the curves $\mathcal{C}$ are straight lines. In Fig. 3.3, we have plotted three characteristics: the slope of these lines is given by the initial condition $u_0(x)$.

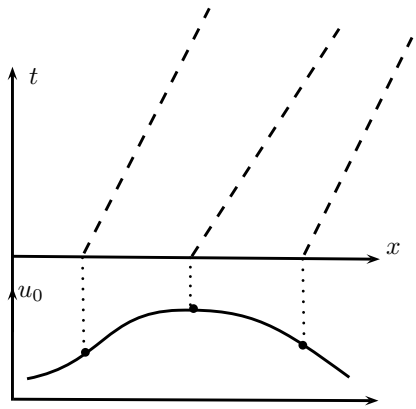


Figure 3.3. *Characteristic curves (dashed lines) for the one-dimensional problem.*

From these transformations, we can deduce that Equations (3.19) and (3.21) are equivalent. Any convection equation can be cast in a characteristic form:

$$\boxed{\frac{\partial}{\partial t}u(x,\ t) + a(u)\frac{\partial}{\partial x}u(x,\ t) = 0 \Leftrightarrow \frac{\mathrm{d}u(x,\ t)}{\mathrm{d}t} = 0 \text{ along straight lines } \mathcal{C} \text{ of equation } \frac{\mathrm{d}x}{\mathrm{d}t} = a(u).} \tag{3.22}$$

When this equation is subject to an initial condition of the form (3.20), the characteristic equation (3.21) can be easily solved. Let us first seek the equation of (straight) characteristics by integrating the differential characteristic equation, knowing that u is constant along the characteristic line:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = a(u) \Rightarrow x - x_0 = a(u)(t - t_0),$$

with the initial condition $t_0 = 0$, $u(x,\ t) = u_0(x)$. We then infer

$$x - x_0 = a(u_0(x_0))t \tag{3.23}$$

is the equation for the (straight) characteristic line emanating from point x_0. Furthermore, we have for $t \geq 0$ $u(x,\ t) = u_0(x_0)$ since u is conserved. Since after equation (3.23), we have: $x_0 = x - a(u_0(x_0))t$, we eventually deduce :

$$\boxed{u(x,\ t) = u_0(x - a(u_0(x_0))t).} \tag{3.24}$$

3.2.2 Shock formation

A feature of hyperbolic equations is that they can propagate an initial discontinuity or generate a discontinuity after a finite time. It is therefore necessary to spend some time on characterizing discontinuities, which here we call *shock*.

We study the formation of a shock for a problem as simple as possible. We consider the convective nonlinear equation:

$$\frac{\partial}{\partial t}u(x,\,t) + \frac{\partial}{\partial x}f[u(x,\,t)] = 0, \tag{3.25}$$

with initial condition $u(x,\,0) = u_0(x)$ and f a given function of u. This equation can be solved simply by the method of characteristics. Indeed, we have seen that a convection equation such as (3.25) can be cast in the following form

$$\frac{\mathrm{d}u}{\mathrm{d}t} = 0 \text{ along curves } \frac{\mathrm{d}x}{\mathrm{d}t} = \lambda(u),$$

We deduce that u is constant along the characteristic curves. So $\mathrm{d}x/\mathrm{d}t = \lambda(u) = c$, with c a constant that can be determined using the initial condition: the characteristics are straight lines with slopes $\lambda(u_0(x_0))$ depending on the initial condition:

$$x = x_0 + \lambda(u_0(x_0))t.$$

Since u is constant along a characteristic curve, we find:

$$u(x,\,t) = u_0(x_0) = u_0\left(x - \lambda(u_0(x_0))t\right)$$

As shown in Fig. 3.4, the characteristic lines can intersect in some cases, particularly when the characteristic velocity decreases (the $x-t$ diagram is in a fact a $t-x$ diagram, this slowdown is reflected in a steepening of the characteristic curves): $\lambda'(u) < 0$. What happens then? When two characteristic curves intersect, this means that potentially, u takes two different values, which is not possible for a continuous solution. The solution becomes discontinuous: a shock is formed.

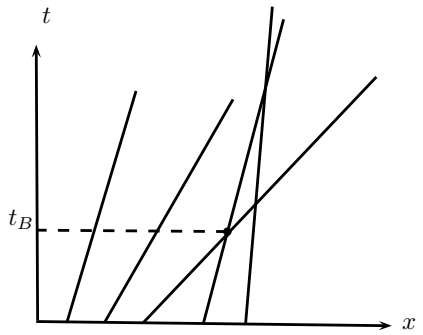


Figure 3.4. *Characteristic curves and shock formation.*

When two characteristic curves interest, the differential u_x becomes infinite (since u takes two values at the same time). We can write u_x as follows

$$u_x = u_0'(x_0)\frac{\partial x_0}{\partial x} = u_0'(x_0)\frac{1}{1 + \lambda'(u_0(x_0))u'(x_0)t} = \frac{u_0'(x_0)}{1 + \partial_x\lambda(x_0)t},$$

where we used the relation: $\lambda'(u_0(x_0))u'(x_0) = \partial_u\lambda\partial_x u = \partial_x\lambda$. The differential u_x tends to infinity when the denominator tends to 0, i.e. at time: $t_b = -1/\lambda'(x_0)$. At the crossing point, u changes its value very fast: a shock is formed. The $s = s(t)$ line in the $x - t$ plane is the shock locus. A necessary condition for shock occurrence is then $t_b > 0$:

$$\boxed{\lambda'(x_0) < 0.}$$

Therefore there is a slower speed characteristic (see Fig. 3.4).

The characteristic curves that are causing the shock form an envelope curve whose implicit equation is given by:

$$x = x_0 + \lambda(u_0(x_0))t \quad \text{et} \quad \lambda'(u_0(x_0)) + 1 = 0. \tag{3.26}$$

After the shock, the solution is multivalued (see Fig. 3.5), which is impossible from a physical standpoint. The multivalued part of the curve is then replaced by a discontinuity positioned so that the lobes of both sides are of equal area.

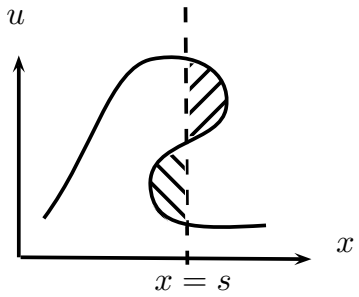


Figure 3.5. *Shock position.*

Generally, we do not attempt to calculate the envelope of characteristic curves, because there is a much simpler method to calculate the trajectory of the shock. Indeed, Eq. (3.25) can also be cast in the integral form:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{x_L}^{x_R} u(x,\ t)\mathrm{d}x = f(u(x_L,\ t)) - (u(x_R,\ t)),$$

where x_L and x_R are abscissa of fixed point of a control volume. If the solution admits a discontinuity in $x = s(t)$ on the interval $[x_L,\ x_R]$, then

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{x_L}^{x_R} u(x,\ t)\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}t}\left(\int_{x_L}^{s} u(x,\ t)\mathrm{d}x + \int_{s}^{x_R} u(x,\ t)\mathrm{d}x\right),$$

That is:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{x_L}^{x_R} u(x,\ t)\mathrm{d}x = \int_{x_L}^{s} \frac{\partial}{\partial t}u(x,\ t)\mathrm{d}x + \int_{s}^{x_R} \frac{\partial}{\partial t}u(x,\ t)\mathrm{d}x + \dot{s}u(x_L,t) - \dot{s}u(x_R,t).$$

Taking the limit $x_R \to s$ and $x_L \to s$, we deduce:

$$\dot{s}[\![u]\!] = [\![f(u)]\!], \tag{3.27}$$

where

$$[\![u]\!] = u^+ - u^- = \lim_{x\to s, x>s} u - \lim_{x\to s, x<s} u,$$

The + and − signs are used to describe what is happening on the right and left, respectively, of the discontinuity at $x = s(t)$.

In conclusion, the short computations that we just made show that if there is a discontinuity at a point $x = s(t)$, then we must have on both sides of $x = s(t)$:

$$\boxed{\dot{s}[\![u]\!] = [\![f(u)]\!]} \tag{3.28}$$

This relationship is called *Rankine-Hugoniot*. It is fundamental in gas dynamics (it is used to calculate the propagation of a supersonic shock wave) and hydraulics (it is used to calculate the propagation of a hydraulic jump).

3.2.3 Riemann problem for one-dimensional equations ($n = 1$)

We call *Riemann problem* an initial-value problem of the following form:

$$\partial_t u + \partial_x[f(u)] = 0,$$

$$u(x,\ 0) = u_0(x) = \begin{cases} u_L & \text{if } x < 0, \\ u_R & \text{if } x > 0, \end{cases}$$

with u_L et u_R two constants. This problem describes how an initially piecewise constant function u, with a discontinuity in $x = 0$ changes over time. This problem is fundamental to solving theoretical problems and to solving hyperbolic equations numerically. In hydraulics, it is also important because the configuration studied corresponds to the rupture of a dam on dry or wet bottom. In the linear case, an initial discontinuity propagates and never disappears; conversely, for a solution to be discontinuous, the initial condition must include a discontinuity. The nonlinear case is somewhat more complex. We shall see that depending on u_R being larger or smaller than u_L, different solutions may be generated. When $f'(u)$ is an increasing function ($f''(u) > 0$) and $u_L < u_r$, the solution initially discontinuous becomes continuous since a continuous wave called *rarefaction wave* allows to link the two initial states and thus reduce the initial discontinuity. Conversely if $u_L > u_r$, the initial discontinuity propagates and the solution remains discontinuous. Recall also that even if the solution is initially continuous, nonlinear equations can generate discontinuities over time (see § 3.2.2). When the function f is complicated, more or less complicated solutions to the Riemann problem may arise.

Linear case

First let us consider the linear case $f(u) = au$, with a a constant. The solution is straightforward:

$$u(x,\ t) = u_0(x - at) = \begin{cases} u_L & \text{if } x - at < 0, \\ u_R & \text{if } x - at > 0. \end{cases}$$

The discontinuity propagates with a speed a.

Nonlinear case

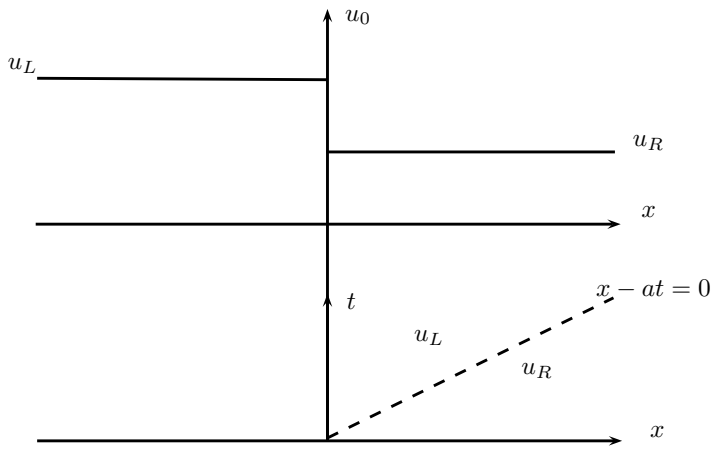


Figure 3.6. *Riemann problem for the linear case.*

Convex flux ($f'' > 0$) In the general case (where $f'' \neq 0$), the Riemann problem is an initial-value problem of the following form:

$$\partial_t u + \partial_x[f(u)] = 0,$$

$$u(x,\ 0) = u_0(x) = \begin{cases} u_L & \text{if } x < 0, \\ u_R & \text{if } x > 0. \end{cases}$$

with u_L and u_R two constants. Assume that $f'' > 0$ (the case of a non-convex flow will not be treated here). We will show that there are two possible solutions:

- a solution called *rarefaction wave* (or simple wave), which is continuous;
- a discontinuous solution which represents the spread of the initial discontinuity (*shock*).

Physically, only one of these solutions is possible and the choice will be dictated by a condition (called *entropy*) depending on the respective value of u_L and u_R. Essentially, the idea is that shocks cause energy dissipation and cannot create energy!

Rarefaction wave. Note first that this equation is invariant under the transformation $x \to \lambda x$ and $t \to \lambda t$. A general solution can be sought in the form $U(\xi)$ with $\xi = x/t$. Substituting this general form into the partial differential equation, we obtain an ordinary differential equation of the form:

$$\left(f'(U(\xi)) - \xi\right) U' = 0.$$

There are two types of solution to this equation:

- *rarefaction wave*: $(f'(U(\xi)) - \xi) = 0$. If $f'' > 0$, then $f'(u_R) > f'(u_L)$; equation $f'(U) = \xi$ admits a single solution when $f'(u_R) > \xi > f'(u_L)$. In this case, u_L is connected to u_R through a *rarefaction wave*: $\xi = f'(U(\xi))$. Inverting f', we find out the desired solution
$$u(x,\ t) = f'^{(-1)}(\xi)\ ;$$
- *constant state* : $U'(\xi) = 0$. This is the trivial solution $u(x,\ t) = \text{cte}$. This solution does not satisfy the initial problem.

The solution reads

$$u(x,\ t) = \begin{cases} u_L & \text{if } \dfrac{x}{t} \leq f'(u_L), \\ f'^{(-1)}(\xi) & \text{si } f'(u_L) \leq \dfrac{x}{t} \leq f'(u_R) \\ u_R & \text{if } \dfrac{x}{t} \geq f'(u_R). \end{cases}$$

Shock wave. We have previously seen that weak solutions (discontinues) to the hyperbolic differential equation (3.25) may exist. Assuming a discontinuity along a line $x = s(t) = \dot{s}t$, we get: $[\![f(u)]\!] = \dot{s}[\![u]\!]$. The solution is then:

$$u(x,\ t) = \begin{cases} u_L & \text{if } x < \dot{s}t, \\ u_R & \text{if } x > \dot{s}t. \end{cases}$$

Then there is formation of a shock wave velocity $\dot{s}$ given by:

$$\dot{s} = \frac{f(u_L) - f(u_R)}{u_L - u_R}.$$

Selection of the physical solution. Two cases may arise (remember that $f'' > 0$). We call $\lambda(u) = f'(u)$ the *characteristic velocity* (see section below), which is the slope of the characteristic curve (straight line) of the problem.

- 1st case: $u_R > u_L$. Since $f'' > 0$, then $\lambda(u_R) > \lambda(u_L)$. At initial time $t = 0$, the two characteristic lines form a fan. Equation $\xi = f'(U(\xi))$ admits a solution over the interval $\lambda(u_R) > \xi > \lambda(u_L)$. See Fig. 3.7;

- 2nd case: $u_R < u_L$. Characteristic lines intersect as of $t = 0$. The shock propagates at rate $\lambda(u_R) < \dot{s} < \lambda(u_L)$. This last condition is called *Lax condition*; it allows to determining whether the shock velocity is physically admissible.

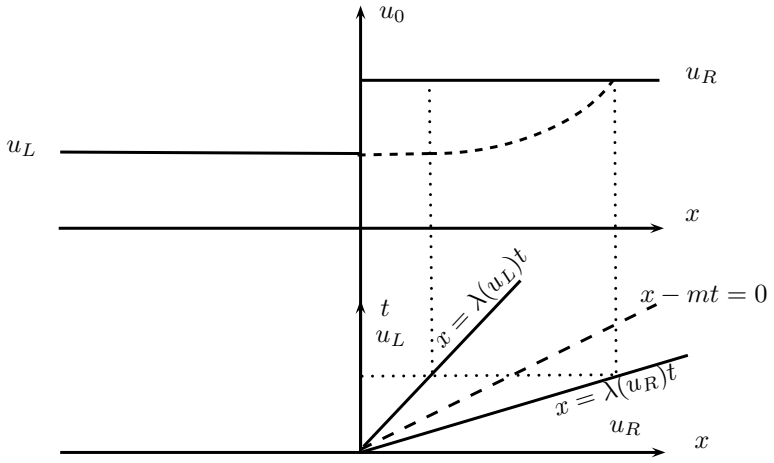


Figure 3.7. *Riemann problem for $u_R > u_L$.*

Non-convex flux For some applications, the flux is not convex. An example is given by the equation of Buckley-Leverett, reflecting changes in water concentration ϕ in a pressure-driven flow of oil in a porous medium:

$$\phi_t + f(\phi)_x = 0,$$

with $f(\phi) = \phi^2(\phi^2 + a(1-\phi)^2)^{-1}$ and a a parameter $(0 < a < 1)$. This fonction has an inflexion point. Contrary to the convex case, for which the solution involves shock and rarefaction waves, the solution is here made up of compose de chocs et *compound wave*) resulting from the superimposition of one shock wave and one rarefaction wave (LeVeque, 2002, see pp. 350–356).

Exercises

Exercise 3.1 We would like to solve the following equation over [0, 1]:

$$\epsilon y'' + y' + y = 0,$$

with initial conditions $y(0) = 1$ and $y(1) = 1$. Show that there is no regular expansion to this equation. Can you find the reason why the regular perturbation techniques does not work here?

Exercise 3.2 (Homework)

The Boussinesq equation is used to calculate the level of an aquifer in the ground; for example, as shown in Figure 3.8, a water flow in a channel, where the water height varies with time causes groundwater flow. For one-dimensional problem, the Boussinesq equation reads

$$\theta \frac{\partial h}{\partial t} = K_s \frac{\partial}{\partial x}\left(h \frac{\partial h}{\partial x}\right),$$

with θ soil porosity, K_s hydraulic conductivity, h water level. The boundary conditions are

$$h(0,\ t) = h_0(t),$$
$$\lim_{x\to\infty} h(x,\ t) = 0.$$

The initial condition is $h(x,\ 0) = 0$ for $x \geq 0$. Reply to the following questions:

- put the Boussinesq equation in a dimensionless form;
- search how to write the similarity solutions when $H_0(t) = At^n$ (i.e., show that solutions can be sought in the form $h = t^\alpha H(\xi)$ with $\xi = xt^{-\beta}$);
- show that the initial differential problem can be transformed into an ordinary differential problem in H;
- we consider the case where the level in the channel is constant $H_0 = A$ ($n = 0$). Show that the governing differential equation in H can be reduced to a quasi-linear equation of first order by making of the following change of variable $z = H/\xi^2$ and $p = H'/\xi$;
- write this equation in the form $\mathrm{d}p/\mathrm{d}z = f(p,\ z)$ with f a function to be determined;
- show that there is front propagation, i.e. a point x_f where $h(x_f) = 0$ and $h = 0$ beyond this point.

Exercise 3.3 Solve the initial value problem:

$$u_x + u_t = 0,$$

for $x \in \mathbb{R}$ and with $u(x,\ 0) = \cos x$.

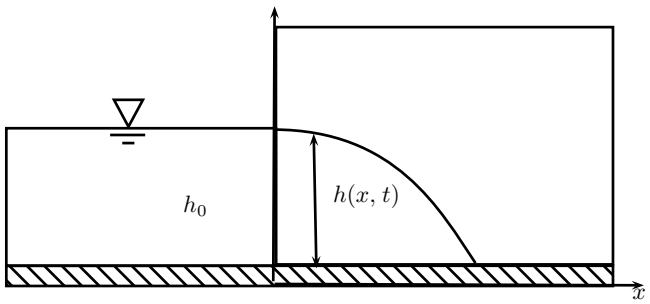


Figure 3.8. *pénétration et propagation d'une nappe aquifère dans un sol.*

Exercise 3.4 Solve analytically, then approximately to order ϵ the following equation

$$\frac{\mathrm{d}y(x)}{\mathrm{d}x} + \epsilon y(x) = 1,$$

with boundary conditions $y(0) = 1$ and where $\epsilon = 0.01$ is a small parameter.

$\rightsquigarrow$ *Réponse* : The analytical solution is

$$y(x) = \frac{e^{-x\epsilon}\left(\epsilon + e^{x\epsilon} - 1\right)}{\epsilon}.$$

We use the following expansion $y = y_0 + \epsilon y_1 + \ldots$. To order ϵ^0, we have to solve $y_0' = 1$ with $y_0(0) = 1$, that is, $y(x) = x + 1$. To order ϵ^1, we have to solve $y_1' + y_0 = 0$ with $y_1(0) = 0$, that is, $y_1(x) = -x - \frac{1}{2}x^2$.

As shown on Fig. 3.9, the deviation between the theoretical and approximate solutions is low, even at order ϵ^0. □

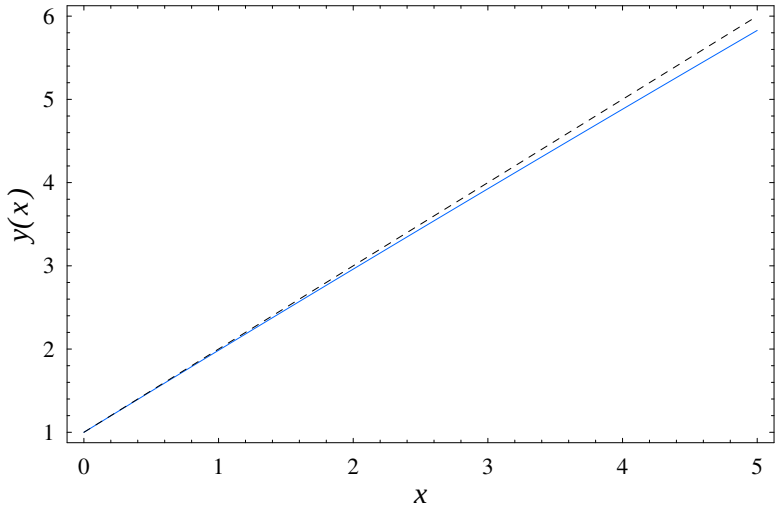


Figure 3.9. *comparison between the theoretical solution (solid line) and approximate solution (to order ϵ^0) $y = x + 1$ (dashed line) over $0 \leq x \leq 5$.*

Exercise 3.5 Solve Huppert's equation, which describes fluid motion over an inclined

plane in the low Reynolds-number limit:

$$\frac{\partial h}{\partial t} + \frac{\rho g h^2 \sin\theta}{\mu}\frac{\partial h}{\partial x} = 0. \tag{3.29}$$

Note that this equation is obtained from the Navier-Stokes equations assuming that the inertial terms are negligible and using the approximation of long wave (Huppert, 1982). The boundary conditions are given in Figure 3.10: it is the release of a finite volume of fluid.

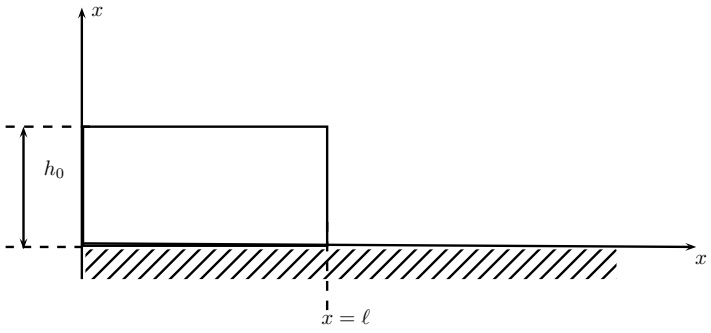


Figure 3.10. *initial flow configuration.*

$\rightsquigarrow$ *Réponse* : This is a nonlinear convection equation of the form: $\partial_t h + c(h)\partial_x h = 0$ with $c(h) = \rho g h^2 \sin\theta/\mu$ or equivalently $\partial_t h + \partial_x f(h) = 0$ with $f(h) = \rho g h^3 \sin\theta/(3\mu)$.

This initial boundary-value problem can be solved straightforwardly. It is indeed a double Riemann problem, a first Riemann problem at $x = 0$ and another one $x = \ell$. We must therefore seek weak solutions (shock) and rarefaction waves that are associated with this equation. For weak solutions with a discontinuity in $x = s(t)$, there is a relation that gives h on both sides of $x = s$

$$\dot{s}[\![h]\!] = [\![f(h)]\!] \tag{3.30}$$

as a function of $\dot{s}$, the shock speed. Rarefaction waves are similarity solutions of the form $\mathcal{H}(\xi)$ with $\xi = x/t$. Here, substituting $h(x, t)$ with $\mathcal{H}(\xi)$ into (3.29), we obtain

$$\mathcal{H}'(-\xi + c(\mathcal{H})) = 0,$$

which implies that we have either $\mathcal{H}' = 0$, or

$$\mathcal{H} = \sqrt{\frac{\mu}{\rho g \sin\theta}}\xi. \tag{3.31}$$

When put in a characteristic form, (3.29) becomes

$$\frac{\mathrm{d}h}{\mathrm{d}t} = 0 \text{ le long de } \frac{\mathrm{d}x}{\mathrm{d}t} = c(h). \tag{3.32}$$

It follows that initially the characteristics are straight lines, whose slope is given by $c(h_0)$, as shown in Fig. 3.11.

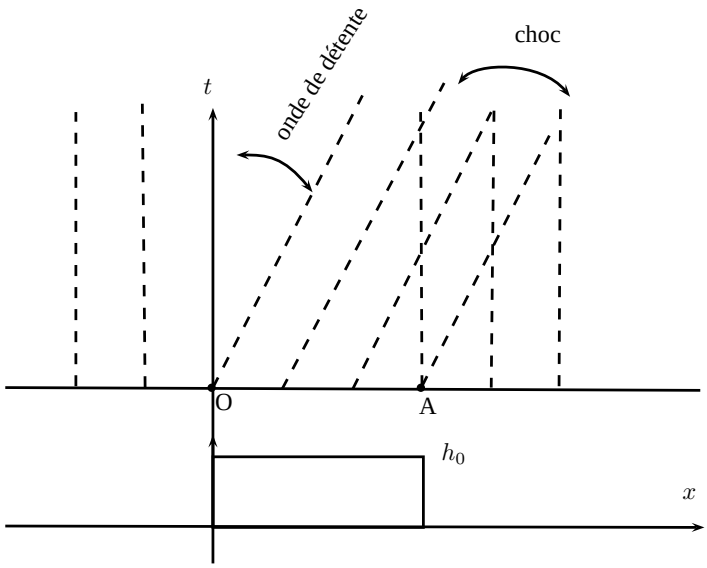


Figure 3.11. *characteristic curves.*

At short times, there is

- a shock on the right. According to (3.28), we have

$$\begin{aligned} \dot{s}_0[\![h]\!] &= [\![f(h)]\!], \\ \dot{s}_0(0-h_0) &= f(0)-f(h_0), \\ \dot{s}_0 &= \frac{f(h_0)}{h_0} = \frac{\rho g h_0^2 \sin\theta}{3\mu} \end{aligned}$$

  whose characteristic curve emanating from point A has the following equation

$$x = \ell + \dot{s}_0 t = \ell + \frac{1}{3}\frac{\rho g h_0^2 \sin\theta}{\mu} t.$$

- a rarefaction wave on the left. Point O is the apex of a fan made up of straight characteristic lines:

$$x = mt,$$

  with m a real parameter varying from 0 to $m_0 = \rho g h_0^2 \sin\theta/\mu$.

The two characteristic lines $x = m_0 t$ and $x = \ell + \dot{s}_0 t$ cross at point B and time t_B

$$t_B = \frac{\ell}{m_0 - \dot{s}_0} = \frac{3}{2}\frac{\mu\ell}{\rho g h_0^2 \sin\theta}.$$

The abscissa of B is

$$x_B = m_0 t_B = \frac{3}{2}\ell.$$

The short-time solution is then ($0 \le t \le t_B$)

$$h(x,\,t) = \sqrt{\frac{\mu}{\rho g \sin\theta}\frac{x}{t}} \text{ for } 0 \le x \le m_0 t, \tag{3.33}$$

$$h(x,\,t) = h_0 \text{ for } m_0 t \le x \le \ell + \dot{s}_0 t \tag{3.34}$$

$$h(x,\,t) = 0 \text{ for } x > s(t) = \ell + \dot{s}_0 t \text{ or } x < 0. \tag{3.35}$$

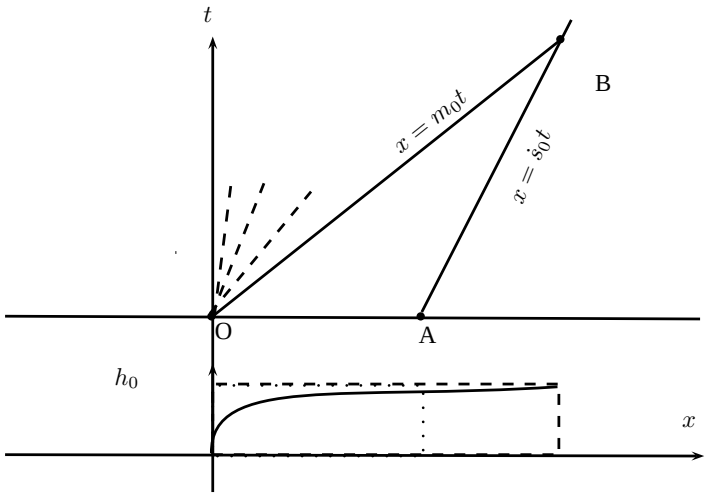


Figure 3.12. *crossing of two characteristics.*

Note that the volume is well conserved over time.

For $t > t_B$, the flow depth on the right of $x = s(t)$ is still 0, but leftwards, it diminishes. According to (3.29), this flow depth is

$$h_s = \sqrt{\frac{\mu}{\rho g \sin\theta}\frac{s}{t}}$$

which leads to the following shock speed

$$\begin{aligned} \dot{s}[\![h]\!] &= [\![f(h)]\!], \\ \dot{s}(0 - h_s) &= f(0) - f(h_s), \\ \dot{s} &= \frac{f(h_s)}{h_s} = \frac{\rho g h_s^2 \sin\theta}{3\mu} = \frac{1}{3}\frac{s}{t}. \end{aligned}$$

Integrating this equation, we find that the characteristic curve associated with the shock is

$$s(t) = x_B \left(\frac{t}{t_B}\right)^{1/3} = \frac{3}{2}\left(\frac{2}{3}\frac{\rho g h_0^2 \ell^2 \sin\theta}{\mu} t\right)^{1/3} = A t^{1/3},$$

with $A = \left(\frac{9}{4}\frac{\rho g h_0^2 \ell^2 \sin\theta}{\mu}\right)^{1/3}$. The long-time solution is then $(t > t_B)$

$$h(x,\, t) = \sqrt{\frac{\mu}{\rho g \sin\theta}\frac{x}{t}} \text{ for } 0 \le x \le A t^{1/3}, \tag{3.36}$$

$$h(x,\, t) = 0 \text{ pour } x > s(t) = A t^{1/3} \text{ for } x < 0. \tag{3.37}$$

Exercise 3.6 The Emden-Fowler equation arises in the context of equilibrium-mass distribution of a cloud of gas with adiabatic exponent

$$\ddot{y} + \frac{2}{x}\dot{y} + y^n = 0,$$

with $n = 5$ (corresponding to an adiabatic exponent 6/5).

- Seek similarity variables (hint: show that the equation is invariant under stretching group).
- Plot the phase portrait after a change of variable (with the similarity variable).
- Try to find an analytical solution.

4 Inverse problems

When solving inverse problem groups different mathematical issues, one tries to invert a relationship in the form $Y = \mathcal{F}(X)$, where Y is a model output or parameter that can measured, X a model input or an internal variable that is not known, and $\mathcal{F}$ a functional. Despite its apparent simplicity and ubiquity in science this remains a complicated issue to which a great deal of research is still devoted. Deterministic methods try to find a solution in the form $X = \mathcal{F}^{-1}(Y)$, but difficult theoretical and numerical issues are usually met: existence, uniqueness, and stability of the solution are rarely ensured in physical problems (Kirsch, 1996). These problems are usually exacerbated when numerical schemes induce discretization errors or when observed data Y are noised. A simple illustration is provided by matrix inversion ($\mathcal{F}$ is then a matrix), when the matrix includes many small off-diagonal terms. Stochastic methods are a second family of techniques, which attempts at finding X by a trial-and-error procedure: guessed values X_n of X are randomly generated, then are tested/selectionned such that $\mathcal{F}(X_n)$ converges towards Y (Ycart, 2002). Since X is determined without making use of $\mathcal{F}^{-1}$, stability is enforced, but other difficulties such as slow convergence can arise. Stochastic methods are particularly recommendable when data are noised and uncertain, which is the case of the available avalanche data.

4.1 Linear inverse problem

We are looking for the solution of the equation:

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b} \tag{4.1}$$

where $\mathbf{A}$ is a $m \times n$ matrix (i.e., it has m rows and n columns), b is a column vector of length m, and $\mathbf{x}$ is the unknown vector. Equation (4.1) is the matrix form of an algebraic linear system, where n unknowns x_j are related by m equations.

When $m = n$, $\mathbf{A}$ is a square matrix; there are as many equations as unknowns and there is a good chance of finding a unique solution to (4.1). When one column is a linear combination of the others, we refer to this situation as *column degeneracy*[1] and the resulting set of equations is *singular*. Numerically this situation can be met when some of the equations are so close to linear dependence that runoff errors make them linearly dependent in the numerical computation.

[1]Row degeneracy entails column degeneracy for square matrices

When $m < n$ (or if $m = n$ but the equations are degenerate), then there are fewer equations than unknowns. In this case there can be either no solution or more than one solution (underdetermined system).

When $m > n$ we are dealing with an overdetermined problem since there are more equations than unknowns. In most cases, there is no solution $\mathbf{x}$ to (**??**, but it happens frequently that we can be satisfied with an approximate solution, i.e. a kind of compromise is sought by determining a vector that almost satisfies all the equations simultaneously, i.e. $\mathbf{x}$ is close to the solution. Here "close to the solution" calls for a proper definition of closeness. This can be achieved for instance by requiring that the $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||^2$ is the smallest as possible. This method is called *the least-square method* and is used to fit a linear curve (line, plane, hyperplane) through a scatter of points.

When $\mathbf{A}$ is square and not singular, methods such as Gaussian elimination or LU decomposition successfully provide the solution $\mathbf{x}$. Otherwise other techniques must be used.

4.1.1 Singular value decomposition for overdetermined systems

Singular value decomposition (SVD) is a technique/algorithm that allows one to decompose any $m \times n$ matrix ($m \geq n$) as the product of an $m \times m$ orthogonal matrix $\mathbf{U}$, an $m \times n$ diagonal-like matrix[2] $\mathbf{W}$ whose elements are positive or zero, and an $n \times n$ orthogonal matrix $\mathbf{V}$:

$$\mathbf{A} = \mathbf{U} \cdot \mathbf{W} \cdot \mathbf{V}^\dagger, \tag{4.2}$$

where $\mathbf{V}^\dagger$ denotes the transpose of $\mathbf{V}$. The matrices $\mathbf{U}$ and $\mathbf{V}$ are orthogonal: $\mathbf{U} \cdot \mathbf{U}^\dagger = \mathbf{1}$ and $\mathbf{V} \cdot \mathbf{V}^\dagger = \mathbf{1}$. The decomposition can always be done and is almost unique, i.e. permutations or linear combinations can be made on the matrices $\mathbf{U}$, $\mathbf{W}$, and $\mathbf{V}$. For a square matrix, if all the diagonal elements of $\mathbf{W}$ are nonzero, than the matrix is not singular and can be inverted:

$$\mathbf{A}^{-1} = \mathbf{V} \cdot \mathrm{diag}[\mathbf{w}_j^{-1}] \cdot \mathbf{U}^\dagger. \tag{4.3}$$

The only thing that can go wrong with this construction is when one of the diagonal elements is zero or very close to zero. Furthermore, when we compute the ratio of the largest element of $\mathbf{W}$ to the smallest entry, a matrix is called *ill-conditioned* if the ratio is too large. If it is infinite, the matrix $\mathbf{A}$ is singular.

If a matrix $\mathbf{A}$ is singular, there is not a single solution. We call *range* of $\mathbf{A}$ the subspace generated by $\mathbf{A}$ (or spanned by the eigenvectors of $\mathbf{A}$). The *rank* of $\mathbf{A}$ is the dimension of this subspace. The complementary subspace is orthogonal to the range and is called the *nullspace*; its dimension is called the *nullity*. By construction we have: rank + nullity $= n$. The nullspace is mapped to zero since for any vector $\mathbf{y}$ in this subspace: $\mathbf{A} \cdot y = 0$. Two cases are to be considered in the equation $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$:

- If $\mathbf{b}$ belongs to the range of $\mathbf{A}$, then there is an infinite number of solutions. Indeed, there is one solution $\mathbf{x}_s$ in the subspace generated by $\mathbf{A}$ and we can add any combinations $\mathbf{r}$ of the basis vectors of the nullspace: $\mathbf{A} \cdot (\mathbf{x}_s + \mathbf{r}) = \mathbf{A} \cdot \mathbf{x}_s = \mathbf{b}$.
- If $\mathbf{b}$ does not belong to the range of $\mathbf{A}$, there is no "exact" solution. However, we can try to find an approximate solution by seeking the solution that does the job in the least-square sense.

[2] $\mathbf{W}$ is diagonal when the matrix $\mathbf{A}$ is square

• In the former case ($\mathbf{b} \in$ Im[$\mathbf{A}$]), we can single out one particular member of this infinite set of solutions by imposing a constraint, e.g. we are looking for the solution with the smallest length. In this case, the technique is to replace $1/w_j$ by zero whenever w_j is zero. Then, we derive:

$$\mathbf{x} = \mathbf{V} \cdot \mathrm{diag}[\tilde{\mathbf{w}}_j^{-1}] \cdot (\mathbf{U}^\dagger \cdot \mathbf{b}), \tag{4.4}$$

where $\tilde{\mathbf{w}}_j^{-1} = 0$ si $\mathbf{w}_j = 0$ and $\tilde{\mathbf{w}}_j^{-1} = \mathbf{w}_j^{-1}$ si $\mathbf{w}_j \neq 0$.

• In the latter case ($\mathbf{b} \notin$ Im[$\mathbf{A}$]), we want to determine a vector $\mathbf{x}$ that minimizes the residual $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||^2$. Indeed, expanding $||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||^2$, then differentiating with respect to $\mathbf{x}$, we obtain:

$$||\mathbf{A} \cdot \mathbf{x} - \mathbf{b}||^2 = (\mathbf{A} \cdot \mathbf{x})^\dagger \cdot \mathbf{A} \cdot \mathbf{x} - 2(\mathbf{A} \cdot \mathbf{x})^\dagger \cdot \mathbf{b} + |\mathbf{b}|^2 = 0$$

$$\frac{\partial}{\partial \mathbf{x}^\dagger} |\mathbf{A} \cdot \mathbf{x} - \mathbf{b}|^2 = 2(\mathbf{A}^\dagger \cdot \mathbf{A}) \cdot \mathbf{x} - 2\mathbf{A}^\dagger \cdot \mathbf{b}$$

The vector that minimizes the residual is solution to the equation: $(\mathbf{A}^\dagger \cdot \mathbf{A}) \cdot \mathbf{x} - \mathbf{A}^\dagger \cdot \mathbf{b}$. Let us now introduce $\mathbf{K} = \mathbf{A}^\dagger \cdot \mathbf{A}$; $\mathbf{K}$ is an $n \times n$ symmetric matrix, thus it has n positive eigenvalues λ_j associated with orthogonal and unit eigenvectors $\mathbf{e}_j$: $\mathbf{K} \cdot \mathbf{e}_j = \lambda_j \mathbf{e}_j$. The vector $\mathbf{x}$ can be projected onto the subspace spanned by $\mathbf{e}_j$:

$$\mathbf{x} = \sum_{j=1}^{n} < \mathbf{x},\ \mathbf{e}_j > \mathbf{e}_j,$$

with $< \mathbf{x},\ \mathbf{e}_j >= \mathbf{x}^\dagger \cdot \mathbf{e}_j$ the scalar product[3]. If $\lambda_j \neq 0$, then $< \mathbf{x},\ \mathbf{e}_j >=< \mathbf{x},\ \lambda_j^{-1}\mathbf{K} \cdot \mathbf{e}_j >$. So we deduce:

$$\mathbf{x} = \lambda_j^{-1} < \mathbf{A} \cdot \mathbf{x},\ \mathbf{A} \cdot \mathbf{e}_j > \mathbf{e}_j,$$

because of the definition of $\mathbf{K}$. Let us introduce $\mathbf{h_j} = \mathbf{A} \cdot \mathbf{e_j}/||\mathbf{A} \cdot \mathbf{e_j}||$. We have ($1 \leq j \leq m$):

$$||\mathbf{A} \cdot \mathbf{e}_j||^2 = (\mathbf{A} \cdot \mathbf{e}_j)^\dagger \cdot (\mathbf{A} \cdot \mathbf{e}_j) = \mathbf{e}_j^\dagger \cdot (\mathbf{A}^\dagger \cdot \mathbf{A}) \cdot \mathbf{e}_j = \mathbf{e}_j^\dagger \cdot (\mathbf{K} \cdot \mathbf{e}_j) = \lambda_j$$

Equivalently :

$$\mathbf{x} = \lambda_j^{-1/2} < \mathbf{A} \cdot \mathbf{x},\ \mathbf{h}_j > \mathbf{e}_j = \lambda_j^{-1/2} < \mathbf{b},\ \mathbf{h}_j > \mathbf{e}_j \tag{4.5}$$

Note that the scalar product $< \mathbf{b},\ \mathbf{h}_j >$ involves vectors of length m while the summation is made for $1 \leq j \leq n$. This is consistent here because $m > n$.

4.1.2 Singular value decomposition for underdetermined systems

If one has m linear equations and n unknowns with $n > m$, there is an $n - m$ dimensional family of solutions. As previously (for overdetermined systems with $\mathbf{b} \notin$ Im[$\mathbf{A}$]) we can apply SVD to underdetermined systems.

The idea is to augment the matrix $\mathbf{A}$ with rows of zeros underneath its nonzero m rows until it is filled up to be square ($n \times n$). Similarly augment with $n - m$ zeros.

[3] $< a,\ b >= a^\dagger \cdot b$

4.2 Fredholm equations: linear techniques

4.2.1 Typology of inverse problems

The problem to be solved

Let us consider the following problem. We have a linear operator K of $\mathcal{L}_2(\mathbb{R})$:

$$K : \mathcal{L}_2(\mathbb{R}) \mapsto \mathcal{L}_2(\mathbb{R})$$
$$f(x) \to g(x) = (Kf)(x).$$

Here we are most interested in linear operators in the form:

$$Kf(x) = \int_a^b f(y)k(x,\ y)\mathrm{d}y \text{ for } c \leq x \leq d. \tag{4.6}$$

where $k(x,\ y)$ is a function of $\mathcal{L}_2(\mathbb{R}^2)$ (or more restrictively of $\mathcal{L}_2([a,\ b])$ called the kernel. The kernel is assumed to be a bounded continuous operator. This equation is called the *Fredholm equation of the first kind*. In practice we have two types of problems:

- We have an analytical expression of g and k and the goal is to recover f.
- We measure a signal $g(x_i)$ at certain times or distances x_i. The goal is to retrieve the function $f(x)$ when the kernel k, the bounds a and b (together with c and d), and the data $g(x_i)$ are known.

We refer to the latter type of problems as the *linear inverse problem with discrete data* as suggested by Bertero *et al.* (1985, 1988). Note that this problem is a bit different from the pure functional problem where the function g is analytically known (and not only a set of sampled data).

Moreover, most of the time, the measurements are inaccurate; so the real problem to solve is:

$$g(x_i) = \int_a^b f(y)k(x_i, y)\mathrm{d}y + \varepsilon_i, \tag{4.7}$$

where ε_i represents noised perturbations, measurement error, uncertainty, etc. The continuous counterpart of this equation can be expressed as follows:

$$\mathrm{d}g(x) = Kf(t)\mathrm{d}t + \mathrm{d}W(t),$$

where $\mathrm{d}W(t)$ is a stochastic process (e.g., a Wiener process).

Presentation of the case study

Here we will examine different techniques and applications to a particular case

$$Kf(x) = \int_a^b f(y)k(x,\ y)\mathrm{d}y \text{ for } c \leq x \leq d. \tag{4.8}$$

where $[a,\ b] = [c,\ d] = [0,\ 1]$ and $K(x, y) = (x - y)^2$; we assume that $f(y) = y^3$, which provides

$$g(x) = \frac{1}{6} - \frac{2\,x}{5} + \frac{x^2}{4}.$$

In the following, we will assume that we know $g(x_i)$ and we will retrieve f.

4.2.2 Collocation method with orthogonal (Legendre) functions

Principle

Assume that we have an orthonormal function basis $P_i(x)$ of $\mathcal{L}^2[a,\ b]$, i.e.:

$$\int_a^b P_i(x)P_j(x)\mathrm{d}x = \delta_{ij}.$$

For example, Hermitte polynomials form an orthogonal basis of $\mathcal{L}^2[\mathbb{R}]$; Legendre polynomials are a orthogonal basis over $[-1,\ 1]$. The normalized Legendre polynomial of order k can be expressed as:

$$P_k(\xi) = \gamma_k \mathrm{d}^k(r^2-1)^k/\mathrm{d}\xi^k,$$

where $\gamma_k = \sqrt{2k+1}/(2^k k!\sqrt{2})$. Over any finite interval $[a,\ b]$, we can define similarly:

$$Q_k = \sqrt{\frac{2}{b-a}} P_k\left(\frac{2}{b-a}(x-a)-1\right)$$

that provides the same properties of P_k, but over the interval $[a,\ b]$. Since Q_k are an orthonormal basis, we can write for any function of $\mathcal{L}^2[a,\ b]$:

$$f(x) = \sum_{k=0}^{\infty} < f,\ Q_j > Q_j(x) = \sum_{k=0}^{\infty} \alpha_j Q_j(x).$$

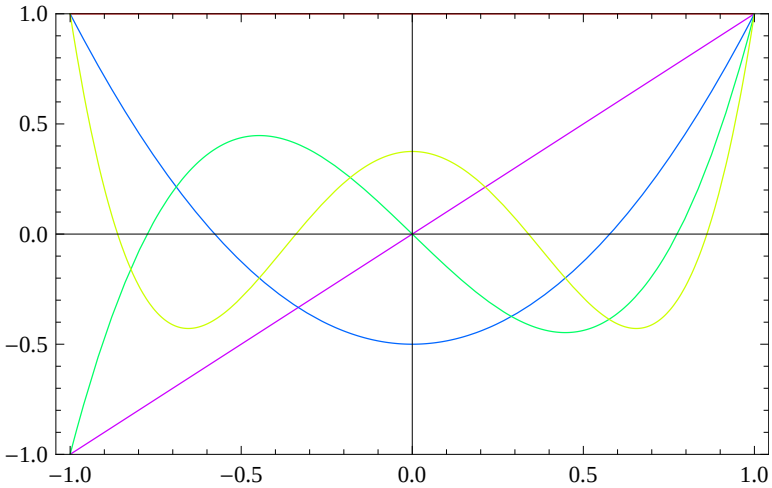


Figure 4.1. *Legendre polynomials $P_i(x)$ of order $i = 0$ to 4.*

In practice, we use a limited number of polynomials:

$$f(x) = \sum_{k=0}^{n_k} \alpha_j Q_j(x),$$

with n_k the degree of the highest polynomial used in the expansion. Note that this truncated expansion is equivalent to projecting f onto the subspace spanned by $(P_i(x))_{1\leq i\leq n_k}$.

In the Fredholm equation, we then deduce:

$$g(x_i) = \int_a^b \alpha_j Q_j(y) K(x_i, y) \mathrm{d}y + \varepsilon_i = A_{ij}\alpha_j + \varepsilon_i, \tag{4.9}$$

where $A_{ij} = \int_a^b Q_j(y)K(x_i, y)\mathrm{d}y$. We finally obtain:

$$\alpha_j = A_{ij}^{-1}(g_i - \varepsilon_i)$$

when $\mathbf{A}$ is invertible. In the event where A is not a square matrix or is not invertible, other techniques such as the singular value decomposition must be used.

Application

We take $x_i = i$ for $1 \leq i \leq 10$. Concerning the number n_k of Legendre polynomials required in the computation, we are restricted by the fact that:

$$A_{ij} = \int_a^b \alpha_j Q_j(y) K(x_i, y) \approx 0$$

for $j \geq 3$. For instance for $n_k = 4$, one obtains:

$$\mathbf{A} = \begin{bmatrix} 0.333333 & -0.5 & 0.166667 & 0. \\ 2.33333 & -1.5 & 0.166667 & 0. \\ 6.33333 & -2.5 & 0.166667 & 0. \\ 12.3333 & -3.5 & 0.166667 & 0. \\ 20.3333 & -4.5 & 0.166667 & 0. \\ 30.3333 & -5.5 & 0.166667 & 0. \\ 42.3333 & -6.5 & 0.166667 & 0. \\ 56.3333 & -7.5 & 0.166667 & 0. \\ 72.3333 & -8.5 & 0.166667 & 0. \\ 90.3333 & -9.5 & 0.166667 & 0. \end{bmatrix}$$

So we can set $n_k = 3$. Using a pseudo-inverse method, we deduce: $\alpha = \{0.25,\ 0.15,\ 0.05\}$.

Obviously one of the major drawbacks of the Legendre polynomials is that we cannot control the accuracy of the computed solution since here, increasing n_k does not improve the final result. Another major inconvenient is the sensitivity of the results when data are noised. Here this comes from the fact that the null space of the matrix $\mathbf{A}$ grows rapidly if we take into account a certain tolerance – specifying how close to zero is considered good enough – in the computations.

For instance, let us consider that measurements are corrupted by noised perturbations $\varepsilon \leftarrow \mathcal{N}(0, \sigma)$ with $\sigma = 0.1$ (a slight perturbation of g). The null space of $\mathbf{A}$ with a tolerance 0 is of dimension 0, but of dimension 3 if we take 0.1 as tolerance, which means that this method is not reliable for the present case. Here, we find $\alpha = \{0.247,\ 0.119,\ -0.273\}$, showing that if the first two coefficients are reasonably well estimated, the last one is far from the value 0.05 found previously.

4.2.3 Collocation method with wavelets

Principle

For this purpose, we decompose f into a wavelet expansion: $f =< f,\ \Psi_\lambda > \Psi_\lambda$, where (see Appendix C) $\lambda = (k,\ l)$ and Ψ_λ refers to ϕ_{kl} or ψ_{kl} depending on the shift and scale indices k

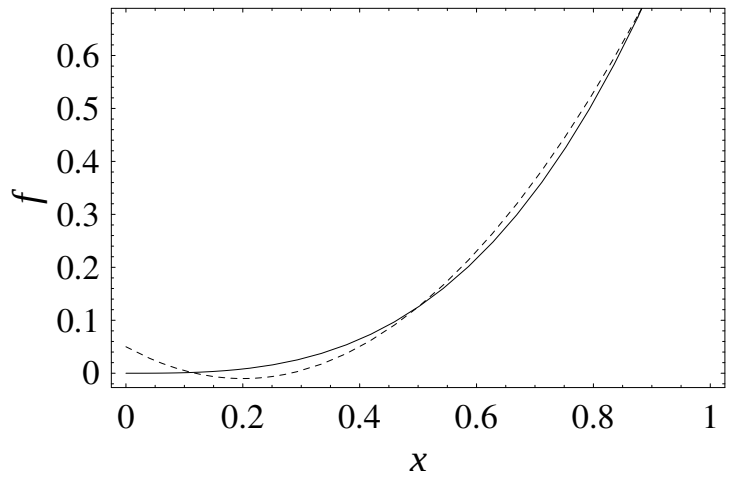


Figure 4.2. *Comparison between the exact solution (solid line) and the computed curve (dashed curve) – no noise.*

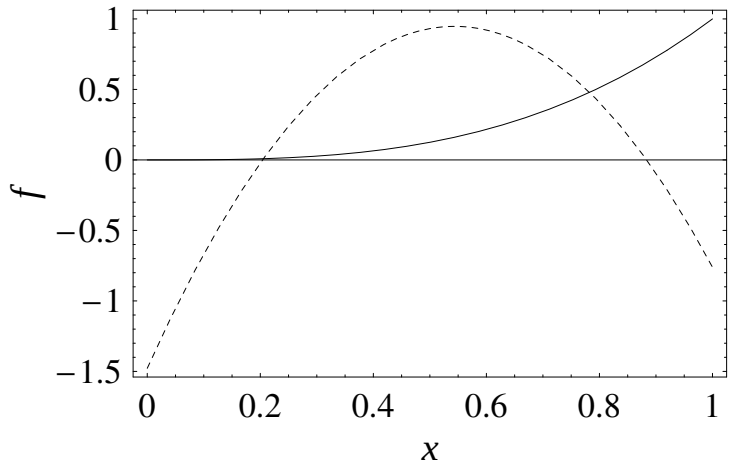


Figure 4.3. *Comparison between the exact solution (solid line) and the computed curve (dashed curve) – noised data ($\sigma = 0.1$).*

and l. We consider $n_k = 2^N$ interpolation points. The inverse problem can then be discretized in the following way:

$$g(x_i) \approx \sum_{\lambda} < f,\ \Psi_\lambda > (K\Psi_\lambda)(x_i)$$

Let us introduce the $n_d \times n_k$ matrix $\mathbf{A}$: $A_{i\lambda} = (K\Psi_\lambda)(x_i)$ and $c_\lambda =< f,\ \Psi_\lambda >$. So we have:

$$g(x_i) \approx A_{i\lambda} c_\lambda$$

We have to solve this linear system, for instance by using the singular-value-decomposition method.

Example

We are going to study the same problem as above:

$$Kf(x) = \int_0^1 k(x,\ y) f(y) \mathrm{d}y,$$

where $k(x,\ y) = (x-y)^2$ and $f(y) = y^3$. Analytically we found $g(x) = \frac{1}{6} - \frac{2x}{5} + \frac{x^2}{4}$. In the inverse problem, we consider $n_d = 10$ data and we are looking for an approximation of f. We have plotted the approximate solution in Fig. 4.4 for two wavelet families: Haar (on the left) and Daubechies D4 (on the right).

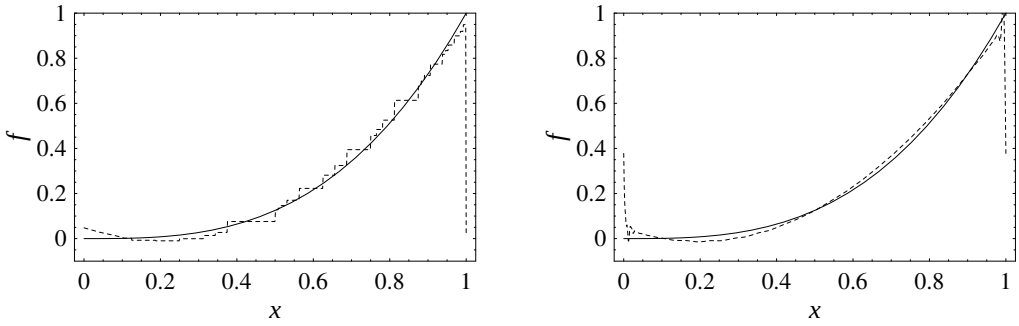


Figure 4.4. *Left: Comparison between the function $f(x) = x^3$ (solid line) and the approximation $\tilde{f}$ solution to the inverse problem when Haar wavelets (left) or Daubechies D4 (right) are used ($n_k = 128$). The matrix* **A** *was inverted by using the SVD method. Eigenvalues lower than a threshold (10^{-3}) were set to zero.*

Fairly good agreement is found. In comparison with other methods, this method exhibits the same deficiencies regarding the sensitivity to noise. A small perturbation of the data leads to substantial differences in the solution.

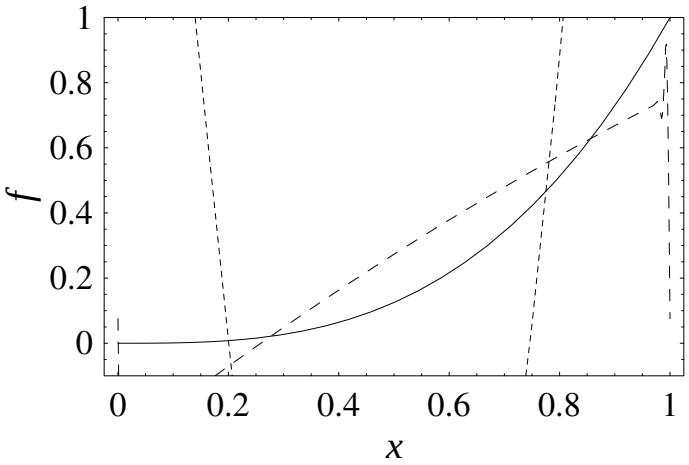


Figure 4.5. *Left: Comparison between the function $f(x) = x^3$ (solid line) and the approximation $\tilde{f}$ solution to the inverse problem when D4 wavelets are used ($n_k = 128$) and when data are noised (Gaussian noise with standard deviation of 0.1): the dashed curve represents the approximate solution obtained by taking Eigenvalues in excess of 10^{-3} (setting to zero those lower than this threshold) and the long dashed curve represents the approximate solution obtained by taking wavelets coefficients in excess of 10^{-2}.*

4.2.4 Galerkin method

Principle

Let us consider the linear operator:

$$K : X \mapsto Y$$

$$f(x) \to g(x) = (Kf)(x) = \int_a^b k(x,\ y)f(y)\mathrm{d}y.$$

We have introduced two function spaces X and Y spanned by a complete set of linearly independent functions ϕ_k and ψ_k respectively ($1 \leq k \leq \infty$), i.e. these sets need not be orthogonal bases. We consider the following approximation of f:

$$\hat{f}(y) = \sum_{i=1}^{n} a_i \phi_i,$$

where the coefficients a_i are determined by setting the residual $r_n = K\hat{f} - g$ orthogonal to $(\psi)_{1 \leq k \leq n}$ with respect to the inner product on Y $< f,\ g >= \int_a^b f(x)g(x)\mathrm{d}x$. So we must have:

$$\sum_{i=1}^{n} a_i < K\phi_i,\ \psi_k >=< g,\ \psi_k >, \quad k = 1, \ldots,\ n.$$

The coefficients a_i are determined by inverting the $n \times n$ matrix $\mathbf{A} =< K\phi_i,\ \psi_k >$. The collocation matrix is cheaper to evaluate since it requires only single integrals in contrast to the double integral needed in Galerkin's method. However, the Galerkin matrix is usually much more sparse than its equivalent in the collocation method (Golberg, 1995). Note also that the number n is directly related to the number of available data and thus it is not possible to improve the accuracy of the solution by increasing the dimension of the space X. Furthermore, in a number of cases of practical interest, the rank of A may be far lower than n, which means that the null space of $\mathbf{A}$ can be of high dimension and a few elements are really available to recover f.

Another major drawback is that in many practical problems, the function g is not known, implying that the inner product $< g,\ \psi_k >$ cannot be computed analytically. Approximate alternative methods could be one of the following:

- using a interpolation function $\hat{g}$ mimicking the behavior of the data $g(x_i)$;
- approximating $< g,\ \psi_k > \approx [g,\ \psi_k] = 1/n \sum_i g(x_i)\psi_k(x_i)$, which is accurate to $1/n$.

In the latter case, we can elaborate a more complete theory by specifying that the kernel maps a space function to a Euclidean space $\mathbb{R}^n$. Let us consider the linear operator:

$$K : X \mapsto \mathbb{R}^n$$

$$f(x) \to g(x_i) = (Kf)(x) = \int_a^b k(x_i,\ y)f(y)\mathrm{d}y \text{ for } i = 1, \ldots,\ n.$$

Let ϕ_k a set of independent functions of X and $\mathbf{e}_i$ a set of vectors spanning $\mathbb{R}^n$; let us call $\tilde{\mathbf{e}}_i = K\phi_i$. Replacing f by $\hat{f}$ gives: $\sum_{i=1}^{n} a_i \tilde{\mathbf{e}}_i = \mathbf{g}$, that is:

$$\sum_{j=1}^{n} a_j < \tilde{\mathbf{e}}_j,\ \mathbf{e}_i >=< \mathbf{g},\ \mathbf{e}_i >, \quad i = 1, \ldots,\ n.$$

Like previously, the coefficients a_i are determined by inverting the $n \times n$ matrix $\mathbf{A} =< \tilde{\mathbf{e}}_j,\ \mathbf{e}_i >$.

Example

The Galerkin method can be used to solve our problem. We need not use an orthogonal basis, which can simplify the problem, especially when we are looking for a full-rank matrix or when we are not interested in exhibiting an orthogonal basis. On Fig. 4.6, we have reported the result obtained by using the purely functional problem and its discrete counterpart. As seen on the figure, there is no difference between the two curves. Note also that the matrices **A** are of rank 3 (which implies that the nullity is 7 since we have 10 data).

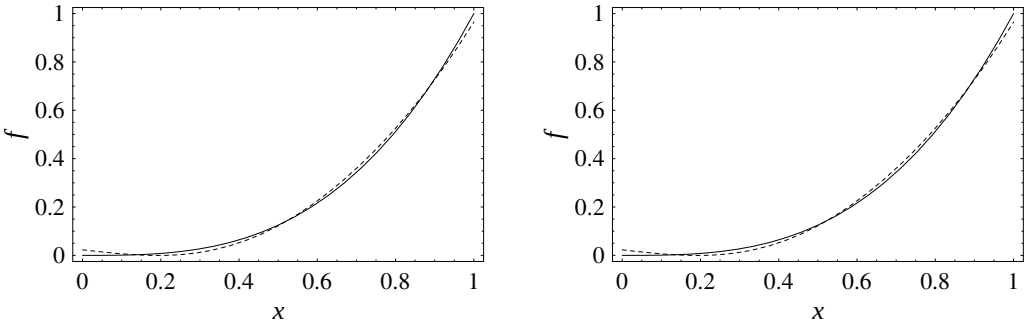


Figure 4.6. *Left: approximate solution (dashed line) versus exact solution (solid line) when Legendre polynomials are used. Left: functional formulation (ψ_i and ϕ_i are Legendre polynomials). Right: discrete formulation (ϕ_i are Legendre polynomials but $\psi_i = \mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$).*

4.2.5 Tikhonov regularization method

Principle

We assume that the experimental database contains n data. To begin with, we consider a mesh Ω with uniform spacing, over which $f(z)$ takes its values. Here, we set $\Omega = [a, b]$. In discretized form Eq. (4.6) becomes:

$$\tilde{g}^C(x_i) = \zeta \sum_{j=1}^{m} \omega_{ij} k(x_i, z_j) f(z_j) \quad 1 \leq i \leq n \tag{4.10}$$

where z_j are the m discretization points of the interval Ω, which are spaced apart by the constant step $\zeta = (b-a)/m$: $z_{j+1} = z_j + j\zeta$, $j = 1, \ldots, m$ with $z_1 = a$. The tilde over g indicates that the quantity is approximated; the true and the approximated quantities, respectively g and $\tilde{g}^C$, differ by an amount that can be estimated on the order of $\zeta^3 \dot{g}$, where $\dot{g}$ denotes the first derivative evaluated at an unknown place in the interval Ω [*Press et al.*, 1992]. To transform the integral term into a sum, we have used a quadrature rule. For a uniform mesh, the simplest scheme is the trapezoidal rule. In this case, the quadrature coefficients are defined as follows: for a given rank i ($1 \leqslant i \leqslant n$), $\omega_{1j} = 1/2$, $\omega_{ij} = 1$, $2 \leq j \leq m-1$, $\omega_{im} = 1/2$. In matrix notation, the discretized equation (4.10) can be written: $\tilde{\mathbf{g}}^C = \zeta \mathbf{W} \cdot \mathbf{f}$, where $\mathbf{W}$ is the matrix defined by $W_{ij} = \omega_{ij} k(x_i, z_j)$, $\tilde{\mathbf{g}}^C = (\tilde{g}_1, \tilde{g}_2 \ldots, \tilde{g}_n)$ the set of computed values, and $\mathbf{f} = (f_1 \ldots, f_m)$ the unknown vector. If we choose $n = m$ and assume that the computed values $\tilde{\mathbf{g}}$ coincide with the measured field data $\tilde{\mathbf{g}}^M = \left(\tilde{g}_1^M, \tilde{g}_2^M \ldots, \tilde{g}_n^M\right)$, we have: $\mathbf{f} = \zeta^{-1}\mathbf{W}^{-1} \cdot \tilde{\mathbf{g}}^C = \zeta^{-1}\mathbf{W}^{-1} \cdot \tilde{\mathbf{g}}^M$ and the inverse problem is easily solved. Otherwise we have m unknown components for n equations.

The simplicity of solving a Fredholm equation is somewhat counterbalanced by the unexpected oscillations around the solution that they produce (Baker, 2000). In order to reduce fluctuations of the discretized solution, it is usually better to proceed by imposing a constraint on the smoothness of the solution. A convenient method is to use *Tikhonov regularization* techniques (Kirsch, 1996). For instance, to ensure a smooth curve, we can impose that the sum of the square of the second derivative at the discretization points, i.e.,

$$\|L_2\mathbf{f}\|^2 = \sum_{j=1}^{m} \left(\frac{\mathrm{d}^2 f}{\mathrm{d}z^2}(z_j) \right)^2 \tag{4.11}$$

is minimal. In Eq. (4.11), L_2 is called second-order derivative operator and $\|\cdot\|$ denotes the vector norm. A finite-difference estimate of the second derivative is $\ddot{f}(z_i) = (f(z_{i+1}) + f(z_{i-1}) - 2f(z_i))/\zeta^2 + o(\zeta^2)$, for $i = 2, 3, \ldots, m-1$. For $i = 1$, assuming symmetry of f leads to: $\ddot{f}(z_1) = (2f(z_2) - 2f(z_1))/\zeta^2 + o(\zeta^2)$. The boundary conditions imposed on f are not too stringent provided that $f \to 0$ when $z \to a$ (resp. $z \to b$); if it is not the case, artifacts may arise from excessive smoothing at the boundaries. We find that the discretized form of Eq. (4.11) is:

$$\sum_{j=1}^{m} \left(\frac{\mathrm{d}^2 f(z_j)}{\mathrm{d}z^2} \right)^2 = G(\mathbf{f}) + o(\zeta^2) \tag{4.12}$$

where G is the quadratic function $G(\mathbf{f}) = (\mathbf{B}\cdot\mathbf{f})^\dagger \cdot (\mathbf{B}\cdot\mathbf{f})$, in which $\mathbf{B}$ denotes the $m \times m$ tridiagonal matrix:

$$\mathbf{B} = \frac{1}{\zeta^2} \begin{bmatrix} -2 & 2 & 0 & \cdots & \cdots & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & -2 & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 1 & -2 & 1 \\ 0 & \cdots & \cdots & \cdots & 0 & 2 & -2 \end{bmatrix} \tag{4.13}$$

$\mathbf{B}$ is the discretized expression of L_2 and therefore is called the second-order difference regularization operator. Other constraints than L_2 can be imposed to ensure the smoothness of the solution. For instance, instead of taking L_2, we can constrain the norm of the solution vector to be as low as possible to avoid overly large fluctuations. In that case, the regularization operator is the identity operator and its discretized expression is $\mathbf{B} = \mathbf{I}_m$, where $\mathbf{I}_m$ is the identity matrix of dimension m.

Determining the unknown vector $\mathbf{f}$ by solving the linear system Eq. (4.6) is equivalent to finding the minimum of the functional:

$$F(\mathbf{f}) = \sum_{i=1}^{n} \left(\frac{\tilde{g}^C(x_i) - \tilde{g}^M(x_i)}{\tilde{g}^M(x_i)} \right)^2,$$

or in matrix form:

$$F(\mathbf{f}) = (\mathbf{1}_n - \mathbf{A}\cdot\mathbf{f})^\dagger \cdot (\mathbf{1}_n - \mathbf{A}\cdot\mathbf{f}),$$

where $A_{ij} = W_{ij}\zeta/\tilde{g}^M(x_j)$ is an $n \times m$ matrix and $\mathbf{1}_n$ is a unity vector of dimension n. If we further assume that g defines a smooth curve, we are looking for a vector that minimizes the functional $H(\mathbf{f}) = F(\mathbf{f}) + \lambda G(\mathbf{f})$, where λ is a free parameter (Lagrangian multiplier). It can be shown that the solution to $\partial H(\mathbf{f})/\partial \mathbf{f} = 0$ has a unique solution, given by (Kirsch, 1996):

$$\mathbf{f} = (\mathbf{A}^\dagger \cdot \mathbf{A} + \lambda \mathbf{B}^\dagger \cdot \mathbf{B})^{-1} \cdot \mathbf{A}^\dagger \cdot \mathbf{1}_n \tag{4.14}$$

Choice of λ

At this point, it should be remembered that the solution actually depends on the Lagrangian multiplier λ, which is an adjustable factor that controls the extent to which the resulting curve is smooth or close to experimental data. The functional F measures something like the agreement of the solution curve to the data, while the functional G reflects the "smoothness" of the curve or the "stability" relative to variations in the data. The better agreement is obtained when a "small" value of λ is chosen, but in this case, the solution may be widely oscillating. In contrast, the best smoothness is produced when using a quite "large" value of λ. In practice, to find a good compromise between agreement and smoothness, the typical idea is to plot the trade-off curve, i.e., $F(\mathbf{f})$ versus $G(\mathbf{f})$ for different values of λ in a log-linear diagram. Generally, the resulting trade-off curve is L-shaped. An appropriate value of λ is then chosen by selecting points in the corner of the L curve (Calvetti *et al.*, 2000). Another method is to select the λ value that minimizes (Wahba, 1990)

$$E(\lambda) = \frac{\|\mathbf{A}\cdot\mathbf{A}-\mathbf{1}_n\|^2}{\left[\text{tr }(\mathbf{I}_m-\mathbf{A}\cdot(\mathbf{A}^\dagger\cdot\mathbf{A}+\lambda\mathbf{B}^\dagger\cdot\mathbf{B})^{-1}\cdot\mathbf{A}^\dagger)\right]^2},$$

with $\mathbf{I}_m$ the identity matrix of dimension m.

Example

We applied the Tikhonov method to the the case study (4.8). Figure 4.7 (a) and (b) shows the approximate solution when, respectively, no regularization is applied and when the Tikhonov technique is used ($\lambda = 10^{-4}$); Fig. 4.7 (c) shows the E curve; a value $\lambda = 10^{-4}$ both minimizes the E function and offers a good trade-off between smoothness and accuracy.

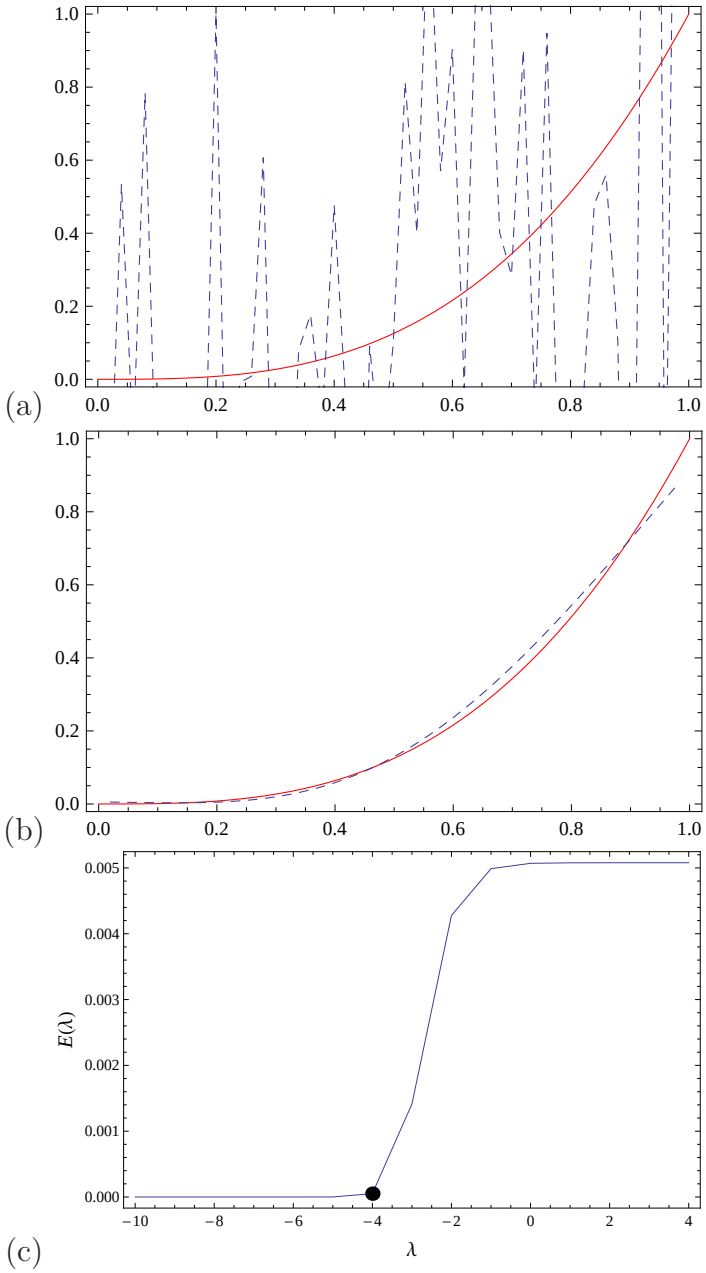


Figure 4.7. *(a) Approximate solution (dashed line) versus exact solution (solid line) when no regularization is applied. (b) approximate solution (dashed line) versus exact solution (solid line) when the Tikhonov technique is used ($\lambda = 10^{-4}$). (c) variation in $E(\lambda)$; the dot represents the value $\lambda = 10^{-4}$.*

4.2.6 Bayesian method and collocation method

Principle

In an uncertain setting, specific attention must be paid to data uncertainty. The Bayesian approach offers an interesting way of both solving inverse problems and estimating the error/uncertainty/credibility related to the computed solution.

The Bayesian approach is based on the Bayes theorem, which derives from the basic axioms of probability theory. Consider two events A and B. The probability of observing A and B is denoted by $P(A, B)$. This probability is given by

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A),$$

where $P(A|B)$ and $P(B|A)$ are conditional probabilities. $P(A|B)$ is the probability of observing A given that B has happened. From the previous relation, we deduce the first Bayes relation

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{4.15}$$

Let us now consider that we have n observations $\mathbf{d} = (d_1, d_2, \ldots, d_n)$. We have a parametric model F for these data; this model depends on k parameters $\mathbf{m} = (m_1, m_2, \ldots, m_k)$. The Bayes relation states that

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}, \tag{4.16}$$

where $p(\mathbf{m}|\mathbf{d})$ is the probability (pdf) of the model parameters, given that we have n data $\mathbf{d}$, i.e. it gives the probability that the model parameters take certain values; it is called the *posterior*. This relation makes it possible to fit a model on data.

Other terms on the right-hand side are important:

- $p(\mathbf{d}|\mathbf{m})$ is the *likelihood*. It represents the probability of observing a data set when the model parameters take certain values $\mathbf{m}$; when $\mathbf{m}$ are close to the right values, then $p(\mathbf{d}|\mathbf{m})$ is high, which means that the data sample $\mathbf{m}$ is likely to be observed. On the opposite, when $\mathbf{m}$ is far from the right values, $p(\mathbf{d}|\mathbf{m})$ is low, giving little hope to observe $\mathbf{d}$.

- $p(\mathbf{d})$ is the probability of observing $\mathbf{d}$. To make sure that $p(\mathbf{m}|\mathbf{d})$ defined by (4.16) is a probability density (i.e., $\int p(\mathbf{m}|\mathbf{d})\mathrm{d}\mathbf{m} = 1$), we set

  $$p(\mathbf{d}) = \int_{\mathbb{R}} p(\mathbf{d}|\mathbf{m})p(\mathbf{m})\mathrm{d}\mathbf{m}.$$

  In analytical calculations, this term causes trouble since there is usually no analytical expression, but in numerical computations, it is not required to know it (it is just a normalization constant). Markov chain algorithms such as the Hastings-Metropolis need not compute $p(\mathbf{d})$.

- $p(\mathbf{m})$ is called the *prior*. It encodes the knowledge we may have of the parameters prior to observation. Because of its speculative character, the notion of prior has been widely debated over the years. The classic viewpoint is that this notion does not make sense since the model parameters are not random variables and moreover probabilities are objective quantities (that can be measured) in the sense that they cannot depend

on *a priori* knowledge. Within the Bayesian perspective, the model parameters are random variables since they are not known accurately; probabilities reflect the degree of belief and therefore are not objective. If we have no expert knowledge, we have to a noninformative prior (e.g., a uniform distribution if the parameters are bounded); the resulting updated probability $p(\mathbf{m}|\mathbf{d})$ will likely remain vague unless the data set is sufficiently dense to refine our knowledge. If we have some reasonably good idea of the values that $\mathbf{m}$ may take, we have to encode this information using appropriate probability distributions (e.g., a normal distribution centered on the desired value); the posterior then gives a measure of our updated state of knowledge.

Bayes' theorem gives us a way of updating our knowledge by following a learning process:

1. we start by encoding our prior knowledge of the model parameters $\mathbf{m}$ into a probability distribution (prior) $p(\mathbf{m})$ that measures the plausibility of our belief or describes some expert opinion;

2. we take measurements and compute the likelihood $L(\mathbf{m} = p(\mathbf{d}|\mathbf{m})$. When measurements are independent, this likelihood is the product of marginal probabilities

$$p(\mathbf{d}|\mathbf{m}) = p(d_1, d_2, \ldots, d_n) = \prod_{i=1}^{n} p(d_i|\mathbf{m}).$$

   Note that this product is usually very small. To avoid roundoff errors in numerical computations, it is better to work with the log-likelihood

$$\ell(\mathbf{m}) = \ln L(\mathbf{m}).$$

   Note also that once the observations have been made, $p(\mathbf{d}|\mathbf{m})$ is considered to be a function of $\mathbf{m}$ solely;

3. in principle, we have to compute the normalization constant $p(\mathbf{d})$, but in practical applications, this is never done;

4. we end up with the posterior $p(\mathbf{m}|\mathbf{d})$ that reflects the updated state of knowledge. Parameters can be selected by, for instance, taking the values that maximize $p(\mathbf{m}|\mathbf{d})$ (modes).

A major issue in the Bayesian approach, which has been a substantial impediment until the advent of computers, lies in the difficulty of computing the posterior. For a long time, the strategy was to select priors such that they lead to analytical computations. With modern computers, there are special computational techniques take generate samples from a given probability distribution. Sampling from any probability distribution P is usually hard, in particular in high-dimensional spaces (i.e., when k is large) because there is no easy way to sample from P without enumerating all possible states. The *Hastings-Metropolis algorithm* offers a convenient and versatile way of generating random samples. A considerable literature has been devoted to this class of algorithm [see (MacKay, 2003; Landau & Binder, 2000), see also (Tanner, 1996; Gilks *et al.*, 1997; Robert, 2001) in the context of Bayesian inference]. We just expose the principles of this method. In the technical literature other MCMC algorithms, including the Gibbs sampler and the simulated annealing method, can be found. They are often more efficient, notably by significantly increasing the rate of convergence, but the Metropolis-Hastings algorithm remains more general and easier to implement.

Hastings-Metropolis algorithm

The basic idea of MCMC algorithms is to introduce a probability distribution from which sampling is straightforward instead of directly sampling from the posterior distribution. We refer to this distribution as the *instrumental distribution* q. We will generate random samples from q and explore the probability space of the posterior distribution; q is then a transition probability which is used to move from a probability state of P_p to another one. If a random value drawn from q is likely to be one of this possible probability state, it may be accepted with a given acceptance rate. Iterating the procedure leads to a sample of values whose empirical probability distribution is close to P_p. In practice, the following steps are performed:

1. Given a current state $\mathbf{X}_n = \mathbf{x}$ draw a candidate value $\mathbf{y}^*$ from the instrumental distribution $q(\mathbf{y}|\mathbf{x})$.
2. Define the acceptance rate r as:

$$r = \begin{cases} \min\left[\frac{P_p(\mathbf{y}^*)q(\mathbf{x}|\mathbf{y}^*)}{P_p(\mathbf{x})q(\mathbf{y}^*|\mathbf{x})},\ 1\right] & \text{if } P_p(\mathbf{x})q(\mathbf{x}|\mathbf{y}^*) > 0 \\ 1 & \text{if } P_p(\mathbf{x})q(\mathbf{x}|\mathbf{y}^*) = 0 \end{cases} \tag{4.17}$$

3. Accept the value $\mathbf{y}_*$ with probability r. In other words, draw a random value u from uniform distribution $\mathcal{U}[0,\ 1]$; if $r > u$, accept $\mathbf{y}_*$ and set $\mathbf{X}_{n+1} = \mathbf{y}^*$ otherwise reject it and set $\mathbf{X}_{n+1} = \mathbf{X}_n$.
4. Repeat the procedure.

The crux of the issue to ensure efficiency of the MCMC simulations lies in the proper selection of an instrumental distribution. Here we have adopted the random-walk version of the Metropolis-Hastings (Robert, 2001), which involves selecting a symmetric probability distribution $q = q(|x-y|)$; note that this choice leads to simplifying the expression of the acceptance rate r in Eq. (4.17): $r = \min[1,\ P_p(\mathbf{y}^*)/P_p(\mathbf{x})]$. Convergence of the empirical distribution of $(\mathbf{X}_n)$ towards P_p is ensured here because of the exponential decrease of the tail of P_p. A common choice is to take a uncorrelated multivariate normal distribution with a tunable covariance matrix ρ

$$q(\cdot|x) :\ y \leftarrow \mathcal{N}(x,\ \rho).$$

The scale matrix ρ must be tuned such that there is a trade-off between the acceptance rate and the ability of the algorithm to fully explore the probability space: if ρ is too large, an extremely large proportion of candidate values will be rejected, leading to a very slow convergence. Conversely, if ρ is too small, the algorithm will nearly accept all the candidate values, but relative motion will be very slow, again leading to inefficiency. A rule of thumb for this algorithm version is to adjust ρ such that the acceptance rate r falls in the range 0.25–0.5. Choosing a uncorrelated distribution q (that is, ρ is a diagonal matrix) makes it possible to adjust the acceptance rate for each component of the vector (X_n).

Formulation

We seek a solution to the following inverse problem with discrete data, which is nothing but a variant of (4.6 when g is known at discrete points

$$g(x_i) = \int_a^b f(y)k(x_i, y)\mathrm{d}y + \varepsilon_i, \tag{4.18}$$

where ε_i represents noised perturbations.

Let us formulate our problem in the following way.

- There are errors introduced by model approximations as well as errors in taking measurements. We assume that the both sources of error can be taken into account using a single error parameter. We then introduce the deviation ε between the computed (g_{comp}) and recorded ($g^{(obs)}$) data: $\varepsilon = g^{(obs)} - g_{comp}$.
- We further assume that ε is a random realization from a normal distribution of mean 0 and unknown variance σ^2: $\varepsilon \leftarrow \mathcal{N}(0, \sigma)$.
- We assume that we have an *a priori* idea of the f dependence on g: $f = F(x; \Theta)$, where Θ denotes the free-parameter set of the functional G. For instance, if a power-law dependence is assumed, then we can express F as a two-parameter function: $f = \alpha x^{-\beta}$, with $\Theta = \{\alpha, \beta\}$. If f is expanded into a truncated series (e.g., Legendre polynomial P_i), then
$$f(x) = \sum_{i=1}^{n_k} a_i P_i(x),$$
and $\Theta = \{a_i\}_{1 \le i \le n_k}$.
- We have a series of n_d measurements: $\mathbf{g} = \left(g_i^{(obs)}\right)$ $(1 \le i \le n_d)$. The Bayes rule allows us to update the parameters Θ using the data and to quantify the uncertainty on Θ:
$$P(\Theta, \sigma|\mathbf{g}, F) = \frac{P(\mathbf{g}|F, \Theta, \sigma)P(\Theta)P(\sigma)}{\int \mathrm{d}\Theta \mathrm{d}\sigma P(\mathbf{g}|F, \Theta, \sigma)P(\Theta)P(\sigma)}, \tag{4.19}$$
where, in the numerator, come up $P(\Theta)$ and $P(\sigma)$ referred to as the probabilities or *priors* of Θ and σ (they are written separately because the variables are assumed to be independent). The quantity
$$P(\mathbf{g}|F, \Theta, \sigma) = \prod_{i=1}^{n_d} P\left((g_i^{(obs)})|F, \Theta, \sigma\right) = \frac{\exp\left[-\sum_{i=1}^{n_d}\left|\frac{g_i^{(obs)} - Kf(x_i)}{\sqrt{2}\sigma}\right|^2\right]}{(\sqrt{2\pi}\sigma)^{n_d}}$$
is called the *likelihood* and refers to the probability of observing the sample $\mathbf{g}$ when the functional F, its parameters Θ, and the standard deviation σ are known. The denominator is a normalizing constant. Recall that the Bayes rule is an updating process, where our knowledge of Θ and σ (knowledge entirely encoded in the prior distributions $P(\Theta)$ and $P(\sigma)$) is updated using the available information $\mathbf{d}$ to provide the *posterior* distribution $P(\Theta, \sigma|\mathbf{g}, F)$.
- In order to obtain an estimate of the best choice for the values of Θ and σ, the idea is to draw random values from the posterior distribution $P(\Theta, \sigma|\mathbf{d}, G)$. If the drawn vector (Θ, σ) is close to the best-fit parameters, then the computed $\mathbf{g}$ vector must be close to the observed values, which leads to a high value of the likelihood $P(\mathbf{d}|F, \Theta, \sigma)$; in other words, the drawn vector is located in a region of high probability. Iterating the process several times, we can obtain an empirical estimate of the posterior probability $P(\mathbf{d}|F, \Theta, \sigma)$. Finally, the best choice of (Θ, σ) can be made by determining the modes or the means of the posterior distribution. The marginal probability density function of

Θ can be estimated by integrating the posterior probability with respect to σ, leading to an estimation of a confidence (credible interval in the Bayesian approach) interval of Θ. Note that the strength of the Bayesian approach lies not only in a proper way of selecting the parameters Θ, but also in a realistic assessment of the uncertainty on Θ and the overall error (combining model and observation errors) ε.

What we need now is to specify the priors $P(\Theta)$ and $P(\sigma)$ (encoding our subjective knowledge of Θ and σ) and the iterative procedure to draw random samples from the posterior distribution $P_p = P(\mathbf{d}|F, \Theta, \sigma)$:

- A common assumption is to consider that the prior of Θ is a multivariate normal distribution of mean Θ_0 and covariance matrix Σ: $P(\Theta) = \mathcal{N}(\Theta_0, \Sigma)$.
- Sampling from the posterior distribution P_p can be done using the efficient and robust Metropolis-Hastings algorithm, which is based essentially on Markov chain sampling and Monte Carlo simulations (MCMC simulations).

Application

We return to the case study (4.8). We will apply two strategies:

- We assume that the standard error σ is known and we are seeking the expansion of f

$$f(y) = \sum_{i=0}^{2} a_i P_i(y)$$

  where P_i is a (normalized) Legendre polynomial. We use the collocation method. For the same reason as previously, we must take $n_k = 3$ polynomials; coefficients associated with higher order polynomials are zero (the kernel being a second-order polynomial). We are using the Hastings-Metropolis algorithm to determine the coefficients a_i.
- we do the same, but we are also looking for the value of the standard error σ, which may make the computations heavier.

For the first strategy, we proceed as follows:

1. We assume that the prior for $\mathbf{a} = (a_i)$ is a multivariate normal distribution

$$\mathbf{a} \leftarrow \mathcal{N}(0, \boldsymbol{\rho})$$

   with mean 0 and covariance matrix $\boldsymbol{\rho}$. The diagonal entries are set to 0.5.

2. the instrumental distribution is a multivariate normal distribution

$$q(\cdot|\mathbf{x}) : \mathbf{y} \leftarrow \mathcal{N}(\mathbf{x}, \boldsymbol{\rho}')$$

   whose correlation matrix is a diagonal matrix, whose entries are set to 10^{-3}. This parameter has been tuned so that the acceptance rate is in the 0.25–0.5 range.

3. We arbitrarily set σ to 0.01.

4. We run the Hastings-Metropolis algorithm with $n = 20\ 000$ iterations.

As shown on Fig. 4.8 with a_0, we can generate samples that explore a reasonably wide range of values around of 0. The histogram shows that the posterior of a_0 is nearly Gaussian, which may be inherited from the prior, but note that the latter was much wider. On Fig. 4.9, we report the mean of the samples generated by the Hastings-Metropolis algorithm as a function of the sample size; we can observe slow convergence towards the right values found in § 4.2.2: $a_0 = 0.25$, $a_1 = 0.15$, and $a_2 = 0.05$.

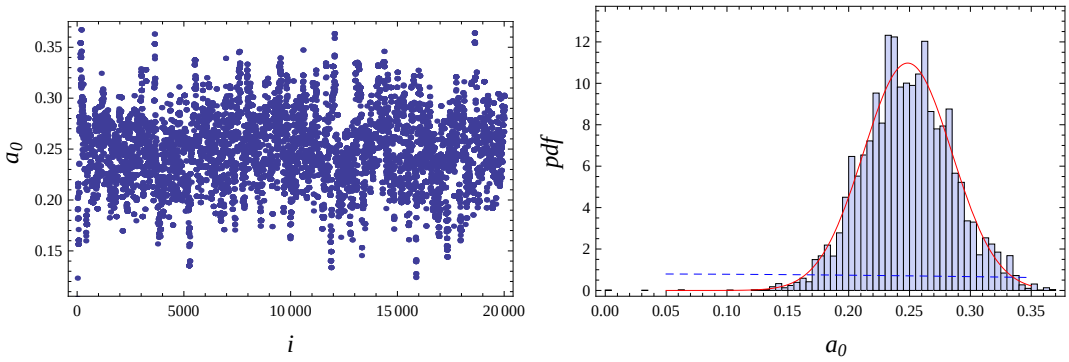


Figure 4.8. *On the left, sample generated by the Hastings-Metropolis and on the right histogram of this sample (computations with $\sigma = 0.01$); the dashed curve is the prior.*

We repeat the same exercise, but by considering that σ is not known. The prior is the uniform distribution over the 0–0.5 range. As instrumental distribution, we take a normal distribution

$$q(\cdot|\sigma) : \; s \leftarrow \mathcal{N}(\sigma, 0.05).$$

As shown on Fig. 4.11, this results in a slow convergence of the coefficients a_i. In fact, after 20 000 iterations, there is no strict convergence, but taking the mean values as an estimate of a_i provides a reasonably good approximate solution as shown by Fig. 4.10.

Although, we do not present results here, it worth noting that the method is as sensitive to noise as the deterministic collocation method. There is no neat advantage in using a Bayesian approach here (except to illustrate it).

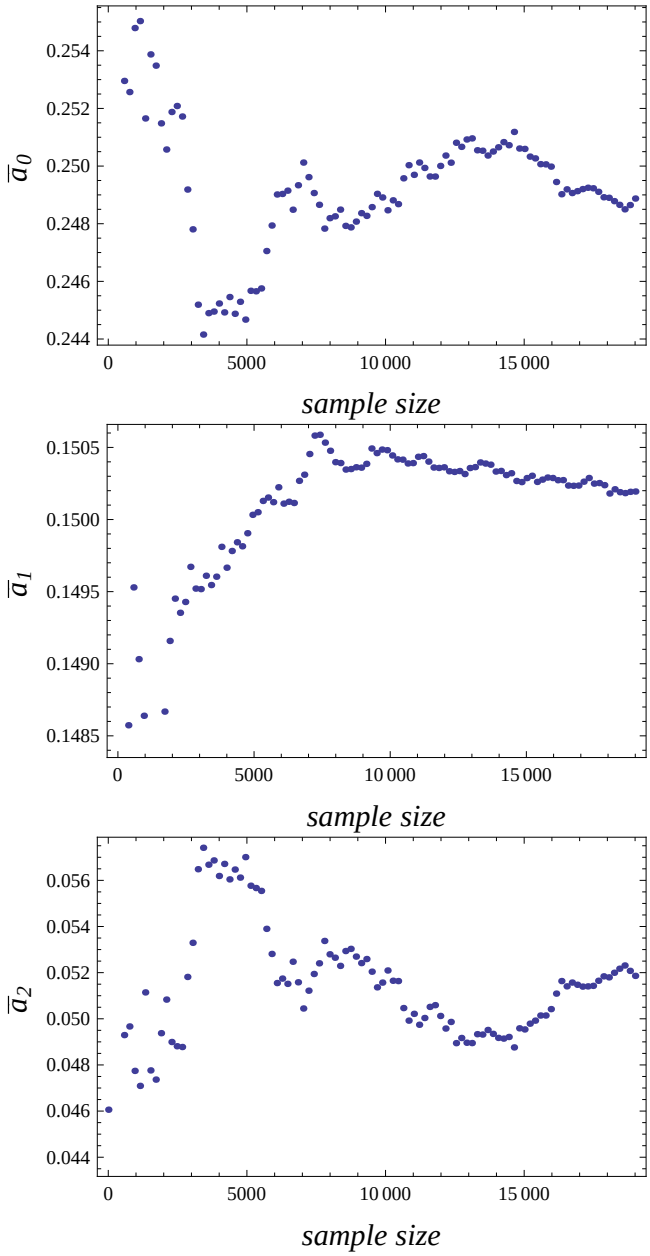


Figure 4.9. *Variations in the mean values of* $a_0 = 0.25$, $a_1 = 0.15$, *and* $a_2 = 0.05$ *depending on the sample size (computations with* $\sigma = 0.01$*).*

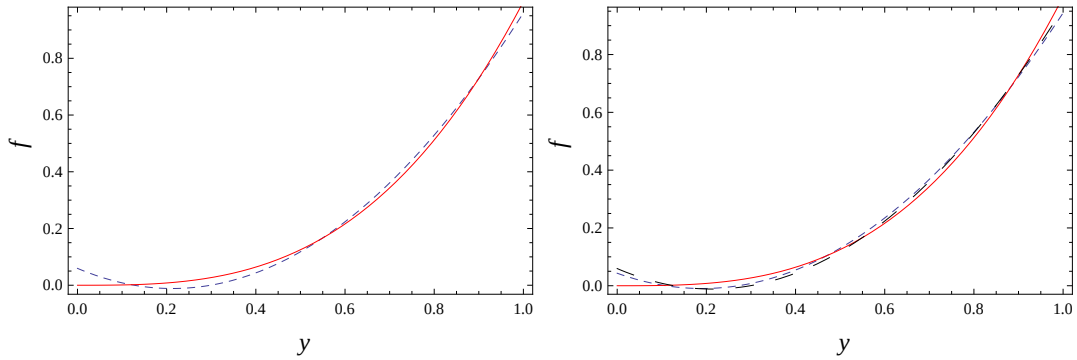


Figure 4.10. *On the left, comparison between the right analytical solution $f(y) = y^3$ (solid line) and the approximate solution (dashed curve) when the standard error is arbitrarily set to $\sigma = 0.1$. On the right, we report the approximate solution when the standard error is a random variable that is used in the Hastings-Metropolis algorithm (the dotted line is the approximate solution corresponding to $\sigma = 0.01$).*

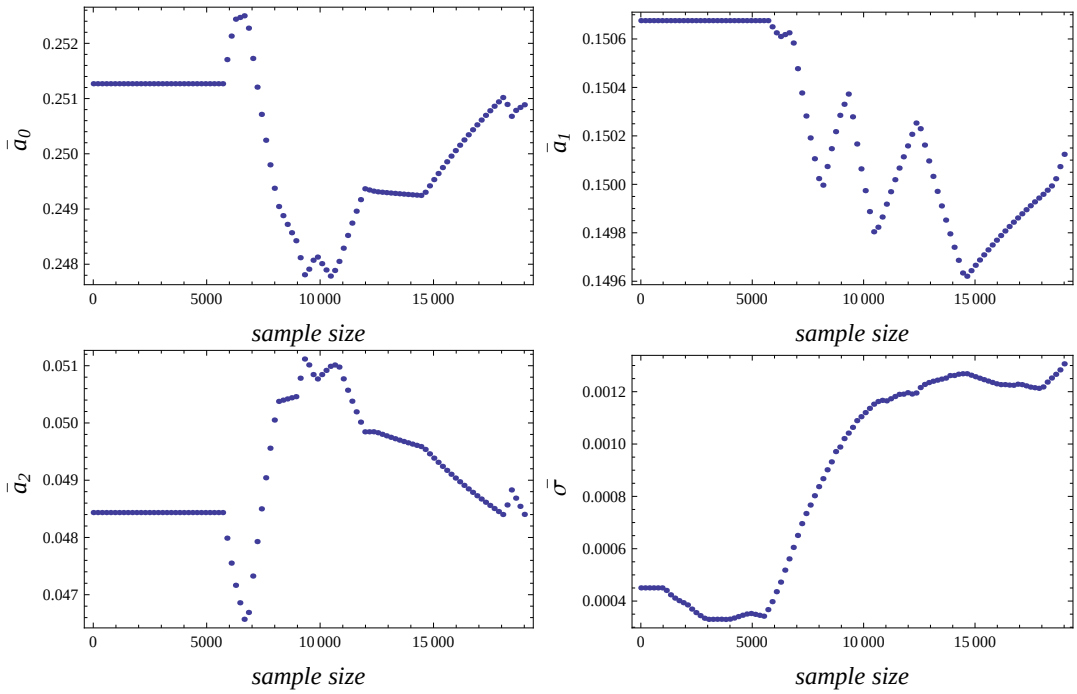


Figure 4.11. *Variations in the mean values of $a_0 = 0.25$, $a_1 = 0.15$, and $a_2 = 0.05$ depending on the sample size (computations with σ considered a random variable).*

4.3 Fredholm equations: nonlinear techniques and refined methods

Let us consider the following problem: we have a linear operator K of $\mathcal{L}_2(\mathbb{R})$:

$$g(x) = (Kf)(x).$$

We have a set of data $g(x_i)$ for $1 \leq i \leq n_d$. The goal is to determine the function f from this set of data. Here we are most interested in linear operators in the form (Fredholm equation of the first kind): $Kf = \int_I k(x,\ y)f(y)\mathrm{d}y$ where $k(x,\ y)$ is a function of $\mathcal{L}_2(\mathbb{R}^2)$ called the kernel and $I = [a,\ b]$ is an interval. We refer to this type of problems as the *linear inverse problem with discrete data* as suggested by Bertero *et al.* (1985, 1988). Note that this problem is a bit different from the pure functional problem where the function g is analytically known (and not only a set of sampled data).

There has been an upsurge in interest in finding sparse representations of integral operators in order to improve algorithm efficiency. The idea is to expand the solution (or more precisely the function to be determined) in a properly chosen basis in which the integral operator is represented by a small number of elements that are larger than some small tolerance λ (sometimes referred to as the drop tolerance); then the remaining elements (those lower than λ) are dropped to provide a sparse representation, i.e. a matrix, whose most elements are zero. It worth noticing that the method used for to solve the integral equation is also important in determining sparseness. In particular, Golberg (1995) demonstrated that Galerkin's method is able to provide sparse matrices, whereas collocation does not.

4.3.1 Regularization by basis functions

The classical approach to solving Fredholm equations has been based on some version of the method of least squares. An alternative is based on the use of adjoint operators (Golberg, 1979; Bertero *et al.*, 1985); the idea is to expand the function to be determined f in a basis formed by the eigenfunctions of the self-adjoint operator K^*K.

Here we will not speak about other methods such as the Backus-Gilbert averaging kernel (O'Sullivan, 1986).

The most popular method is the so-called *singular value decomposition* (SVD) that we shall outline below. It is very attractive because of its efficiency, its algorithm simplicity and versatility, and its optimality in many cases. However, it suffers from two limitations:

- in the function formulation of the problem, the eigenfunctions can be difficult to determine or to handle with. Note that this limitation does not hold in the discrete case, as shown below;

- the eigenfunctions describe the action of f but may be unappropriate to provide an accurate decomposition/description of f (i.e. with a small number of parameters).

Principle

Let us introduce the discrete operator K_n which maps $\mathcal{L}_2(\mathbb{R})$ to $\mathbb{R}^n$, i.e. square integrable functions f to vectors of length n $\mathbf{y}$:

$$K_n : f(x) \to y_i = (Kf)(x_i) = \int_I k(x_i, z) f(z) \mathrm{d}z$$

We can define the adjoint operator which maps $\mathbb{R}^n$ to $\mathcal{L}_2(\mathbb{R})$:

$$K_n^* : \mathbf{y} = (y_i) \to h(z) = y_i k(x_i, z)$$

Indeed, let us compute:

$$[Kf,\ \mathbf{y}] = \sum_i y_i \int_I k(x_i,\ z) f(z) \mathrm{d}z = \int_I f(z) \mathrm{d}z \sum_i y_i k(x_i,\ z) =< f,\ \sum_i y_i k(x_i, \cdot) >,$$

This demonstrates that K_n^* is the adjoint operator of K_n. We can now define a basis of singular functions u_i that are the normalized eigenfunctions of $K_n^* K_n$: $K_n^* K_n u_i = \delta_i^2 u_i$. $L_n = K_n^* K_n$ is a self-adjoint operator that is finite-rank and non-negative. It is an integral operator whose kernel is given by:

$$L_n[f](y) = \int_I \sum_{i=1}^{n} k(x_i,\ y) k(x_i,\ z) f(z) \mathrm{d}z.$$

Similarly $M_n = K_n K_n^*$ is a self-adjoint non-negative operator of the finite space $\mathbb{R}^n$. It is called the *Gram matrix*. Therefore it can be characterized by a $n \times n$ symmetric matrix that has n positive eigenvalues:

$$M_{ij} = \int_I k(x_i,\ z) k(x_j,\ z) \mathrm{d}z.$$

We order the eigenvalues in such a way they form a non-decreasing sequence: $\delta_1^2 \geq \delta_2^2 \geq \ldots \geq \delta_N^2$. Note that L_n and M_n have the same eigenvalues. We denote by u_k the normalized eigenvalues of M_n; the functions u_k are orthogonal (and orthonormal) in $\mathcal{L}^2(\mathbb{R})$. Similarly we define the eigenvectors $\mathbf{v}_k$ of the matrix $\mathbf{M}$ (characterizing the operator M_n). It is possible to select the eigenfunctions and eigenvectors in such a way that:

$$\boxed{K_n u_k = \delta_k \mathbf{v}_k \ \text{ and } \ K_n^* \mathbf{v}_k = \delta_k u_k.} \tag{4.20}$$

Indeed, starting from: $K_n^* K_n u_i = \delta_i^2 u_i$, applying the operator K_n, we have: $(K_n K_n^*) K_n u_i = \delta_i^2 K_n u_i$, showing that $K_n u_i$ is an eigenvector of L_n associated with δ_i^2, therefore it must be proportional to $\mathbf{v}_i$ and it is possible to adjust the proportionality factor such that: $K_n u_k = \delta_k \mathbf{v_k}$.

We are looking for the approximate solution $\tilde{f}$ solution to $K_n f = \mathbf{g}$, with $\mathbf{g}$ the set $(g(x_1), \ldots, g(x_n))$. So we have also: $K_n^* K_n \tilde{f} = K_n^* \mathbf{g}$. Since $\tilde{f}$ can be represented by a linear combination of the singular functions u_i that form an orthonormal basis ($\tilde{f} =< \tilde{f},\ u_i > u_i$), we obtain:

$$\tilde{f} =< \tilde{f},\ u_i > u_i =< \tilde{f},\ K_n^* \mathbf{v}_i > \frac{u_i}{\delta_i} = [K_n \tilde{f},\ \mathbf{v}_i] \frac{u_i}{\delta_i},$$

and we finally deduce (Bertero *et al.*, 1985):

$$\boxed{\tilde{f} \approx [\mathbf{g},\ \mathbf{v}_i] \frac{u_i}{\delta_i},} \tag{4.21}$$

Example

Let us take again the previous example. First compute the matrix $\mathbf{M}$. Here we take: $x_i = i$ for $i = 1 \ldots 10$. We find $\mathbf{M} = \int k(x_i,\ y)k(x_j,\ y)\mathrm{d}y$. There are 3 non-zero eigenvalues: 19993.2, 6.76, and 0.0059. We deduce the eigenfunctions (see 4.12) and then can compute the approximate solution. Note that this method is also sensitive to noise because of the tremendous ratio between the largest and smallest eigenvalues (ratio of 3.35×10^6). Here this can be explained by the fact that the Gram matrix $\mathbf{M}$ has widely varying coefficients. Even by increasing the data number n_d, we that $\mathbf{M}$ has only 3 non-zero eigenvalues, which shows that the range of $\mathbf{M}$ is of dimension 3 and the null space of dimensions $n_d - 3$. Here typically there are only 2 eigenvalues such that $\delta_i^2 \gg \sigma^2$ where σ denotes the noise variance. So, in the present case, there is little hope to recover f when data are noised.

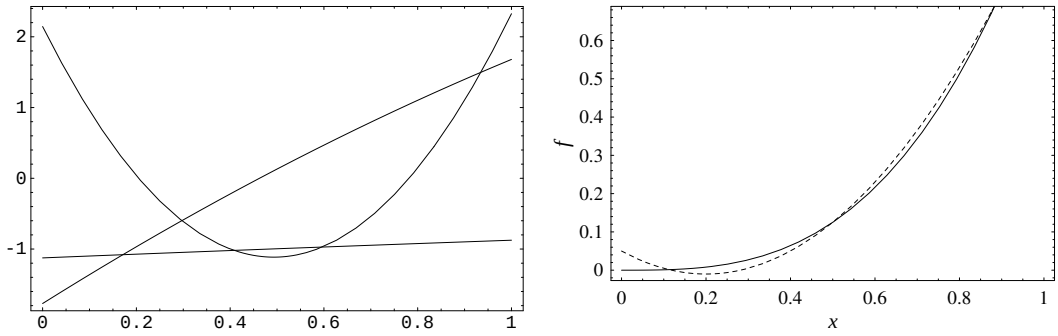


Figure 4.12. *Left: Eigenfunctions of K_n^*. Right: approximate solution (dashed line) versus exact solution (solid line).*

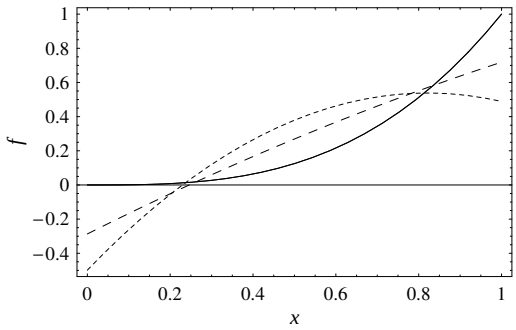


Figure 4.13. *Approximate solution (dashed line) versus exact solution (solid line) when data are corrupted by noise (Gaussian noise with standard deviation of 0.1). The long-dashed curve is the solution obtained by removing the third eigenvalues (0.0059) that is far smaller than the first.*

Following Donoho (1995), this can be seen as following: we observe $\mathbf{g}$ modelled as $\mathbf{g} = K_n f + \sigma\epsilon$, where $\sigma\epsilon$ is a noise of strength σ. The scalar product with $\mathbf{v}_i$ provides

$$[\mathbf{g},\ \mathbf{v}_i] = [K_n f + \sigma\epsilon,\ \mathbf{v}_i] = [K_n f,\ \mathbf{v}_i] + \sigma[\epsilon,\ \mathbf{v}_i] = [f,\ K_n^*\mathbf{v}_i] + \sigma =< f,\ u_i > \delta_i + \sigma$$

We deduce that

$$\tilde{f} = \sum_i < f,\ u_i > u_i + \frac{\sigma}{\delta_i} u_i$$

and the mean-square error is then

$$\mathrm{MSE}[\tilde{f}] = ||f - \tilde{f}||^2 = \left| |f - \left(\sum_i < f,\ u_i > u_i + \frac{\sigma}{\delta_i} u_i \right) \right|^2$$

$$\mathrm{MSE}[\tilde{f}] \leq \left| f - \sum_i < f,\ u_i > u_i \right|^2 + \left| \frac{\sigma}{\delta_i} \right|^2$$

4.3.2 Wavelet-vaguelette decomposition

Wavelet bases offer efficient and sparse representations for functions in a wide range of function spaces; they have demonstrated certain success in nonparametric function estimation in terms of spatial adaptivity and asymptotic optimality. Donoho (1995) proposed the *wavelet-vaguelette* decomposition (WVD) that involves expanding the function f into a wavelet series resulting in a vaguelette series for the function $g = Kf$. Abramovich & Silverman (1998) proposed a variant, the vaguelette-wavelet decomposition. In both approaches, the wavelet coefficients are thresholded to reduce noise.

Not all linear operators can be decomposed using the WVD approach. Donoho (1995) demonstrated that this decomposition is possible for certain classes of operator, including homogenous operators. According to Donoho (1995), an operator is homogenous with respect to dilation. Let $D_a[f](t) = f(at)$; the operator is homogenous of degree α if the operators intertwine: $KD_a = a^\alpha D_a K$. The vaguelettes are functions defined as the images of the wavelet ψ:

$$e_{ij} = K_{ij}$$

If the operator K is homogenous and intertwine with shift operators, then e_{ij} are also wavelets.

The method has been applied to a number of problems: Wicksell problem (size distribution determination) (Champier & Grammont, 2002)

Principle

Let us consider the wavelet decomposition of f: $f(x) = \sum_{i,j} a_{ij}\psi_{ij}$ where $a_{ij} =< \psi_{ij},\ f >$. By applying K we obtain the vaguelette decomposition with the same coefficients:

$$g(x) = \sum_{i,j} a_{ij} e_{ij}$$

So the idea is to find the vaguelette decomposition of g, then to recover f. The problem is that e_{ij} are not an orthonormal basis, we thus need use a dual basis $\tilde{e}_{ij}$ such that $< e_{ij},\ \tilde{e}_{kl} >= \delta_{ik}\delta_{jl}$.

As in the previous section we introduce the operator K_n that maps $\mathcal{L}(\mathbb{R})$ to $\mathbb{R}^n$ and its adjoint operator K_n^*:

$$K_n : f(x) \to y_i = (Kf)(x_i) = \int_I k(x_i, z) f(z) \mathrm{d}z$$

$$K_n^* : \mathbf{y} = (y_i)_{1 \leq i \leq n} \to h(z) = \sum_{i=1}^{n} y_i k(x_i, z)$$

As a consequence, $\mathbf{e}_{ij} = K_n\psi_{ij}$ is a vector in $\mathbb{R}^n$ (and the same for the dual basis). From the relation

$$[\mathbf{e}_{ij},\ \tilde{\mathbf{e}}_{kl}] = [K_n\psi_{ij},\ \tilde{\mathbf{e}}_{kl}] =< \psi_{ij},\ K_n^*\tilde{\mathbf{e}}_{kl} >= \delta_{ik}\delta_{jl},$$

we deduce that ψ_{ij} and $K_n^*\tilde{\mathbf{e}}_{kl}$ are parallel. In other words, there exists a factor β_{ij} such that $\beta_{ij}\psi_{ij} = K_n^*\tilde{\mathbf{e}}_{ij}$. It follows that:

$$a_{ij} =< f,\ \psi_{ij} >=< f,\ \beta_{ij}^{-1}K_n^*\tilde{\mathbf{e}}_{ij} >= \beta_{ij}^{-1}[K_nf,\ \tilde{\mathbf{e}}_{ij}].$$

An estimate of a_{ij} is obtained by replacing K_nf by $\mathbf{g}$:

$$\boxed{a_{ij} =< f,\ \psi_{ij} >\approx \beta_{ij}^{-1}[\mathbf{g},\ \tilde{\mathbf{e}}_{ij}].} \tag{4.22}$$

Note that the factors β_{ij} can be found by

$$\beta_{ij}\psi_{ij} = K_n^*\tilde{\mathbf{e}}_{ij} \Rightarrow \beta_{ij}\mathbf{e}_{ij} = K_nK_n^*\tilde{\mathbf{e}}_{ij}$$

$$\beta_{ij} = \frac{K_nK_n^*\tilde{\mathbf{e}}_{ij}\cdot\mathbf{e}_{ij}}{||\mathbf{e}_{ij}||^2}.$$

A nonlinear thresholding procedure can then be applied to the coefficients before wavelet inverse transform. A trick is to normalize the coefficient $a_{ij} = \beta_{ij}^{-1}[K_nf,\ \tilde{\mathbf{e}}_{ij}]$ by $||\tilde{\mathbf{e}}_{ij}||$ so that the rescaled coefficients $\hat{a}_{ij} = a_{ij}||\tilde{\mathbf{e}}_{ij}||$ have the same variance. Extraction of the important coefficients is based on the idea that the large values $|\hat{a}_{ij}|$ contribute to real signal, whereas the small values carry noise. Applying a soft thresholding procedure

$$\Delta_\lambda(x) = \text{sign}(x)(|x| - \lambda)_+$$

or a hard threshold function

$$\Delta_\lambda(x) = \begin{cases} x \text{ if } |x| > \lambda, \\ 0 \text{ otherwise,} \end{cases}$$

where λ denotes a given threshold. Mapping the thresholded coefficients back into the wavelet expansion in the original space yields the resulting approximation:

$$\hat{f} = \sum_i\sum_j ||\tilde{\mathbf{e}}_{ij}||\Delta_\lambda(\hat{a}_{ij})\psi_{ij}.$$

Donoho (1995) showed that if the threshold λ is properly chosen level by level, then the WVD converges quickly and surely toward the "better" solution. This convergence is described in terms of the minimax rate. The SURE methods provides a method for optimally choosing λ (Donoho & Johnstone, 1998).

Example

We use the same example as previously. We report the solutions for two bases (Daubechies D4 and D8) on Fig. 4.14.

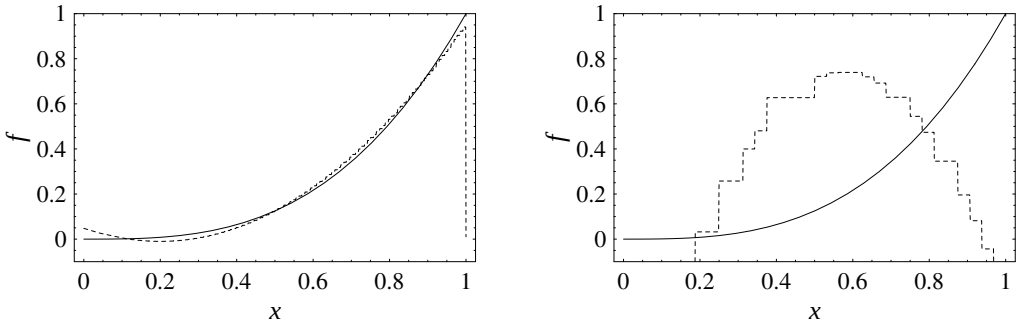


Figure 4.14. *Left: approximate solution (dashed line) versus exact solution (solid line) for Harr wavelet; use $n_k = 128$ functions. On the right, the same exercise was repeated when data are contaminated by noise ($\sigma = 0.1$), note that we use $n_k = 32$.*

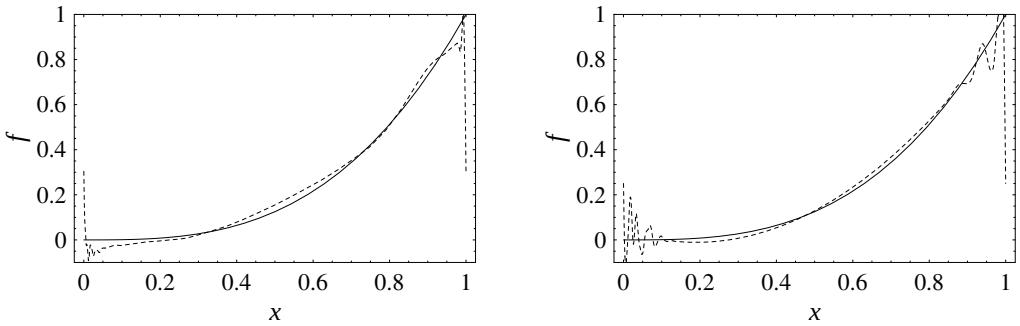


Figure 4.15. *Left: Approximate solution (dashed line) versus exact solution (solid line) for the Daubechies wavelet D4 ($n_k = 128$). Right: approximate solution (dashed line) versus exact solution (solid line) for the Daubechies wavelet D8 ($n_k = 128$).*

4.3.3 The approximate inverse method

Principle

This method was devised by Louis (Louis & Maass, 1990; Louis, 1996, 1999). We start from the adjoint problem:

$$K_n^* K_n f = K_n^* \mathbf{g},$$

with the same notations as those adopted for the SVD decomposition. Here we further assume that the approximate solution $\tilde{f}_n$ belongs to the orthogonal complement of the null space of K_n. This means that we can write:

$$f = \tilde{f}_n + h,$$

where $h \in \mathcal{N}[K_n]$ $(K_n h = 0)$. Said differently: $\mathbb{R} = \mathcal{N}[K_n] + (\mathcal{N}[K_n])^{\perp}$.

We introduce a family of smooth functions, the *mollifiers*, depending on a scalar parameter λ: $e_\lambda(x,\ y)$. A typical example is the hat profile: $e_\lambda(x,\ y) = \chi_{[\lambda,\ \lambda]}(x - y)/(2\lambda)$, where χ is the characteristic function of the interval $[\lambda,\ \lambda]$. A mollifier defines the kernel of a smooth operator E_λ:

$$E_\lambda : \mathcal{L}(\mathbb{R}) \to \mathcal{L}(\mathbb{R})$$

$$E_\lambda : h \to \int_R e_\lambda(x,\ y) h(y) \mathrm{d}y$$

We define a smoothed solution f_λ by setting $f_\lambda = E_\lambda \tilde{f}_n$. This function can be determined by the help of a *reconstruction kernel* $\mathbf{v}_\lambda(x)$, which is the solution of:

$$K_n^* \mathbf{v}_\lambda(x) = e_\lambda(x, \cdot) \tag{4.23}$$

Note the specific structure of this kernel since it is a vector depending on a variable x. We have:

$$\boxed{f_\lambda(x) = [\mathbf{g}, \mathbf{v}_\lambda(x)]} \tag{4.24}$$

In practice, the solution to equation to equation (4.23) may not exist or be to too difficult to determine. In this case, we can content ourselves with an approximate solution that is the solution to the least square problem. For a given (dummy) variable x, we look for the solution of

$$K_n K_n^* \mathbf{v}_\lambda(x) = K_n e_\lambda(x, \cdot).$$

This means that:

- (i) we compute the vector $K_n e_\lambda(x, \cdot)$ for a given x,
- (ii) we compute the matrix $K_n K_n^*$ (that is independent of data),
- (iii) we derive $\mathbf{v}_\lambda(x)$ by computing $\mathbf{v}_\lambda(x) = (K_n K_n^*)^{-1} K_n e_\lambda(x, \cdot)$.

Exemple

Here we modify the study case because the matrix $\mathbf{M} = K_n K_n^*$ is ill-conditioned and singular, thus cannot be inverted. We consider the following problem:

$$Kf = \int_0^x y f(y) \mathrm{d}y = \int_0^1 y H(x-y) f(y) \mathrm{d}y,$$

where H is the Heaviside function. When $g(x) = 2x^{5/2}/5$, the solution to this problem is $f(x) = \sqrt{x}$. We consider the hat-top mollifier $e_\lambda(x, y) = \chi_{[\lambda, \lambda]}(x-y)/(2\lambda)$. As seen on Fig. 4.16, the choice of λ has significant impact on the smoothness of the solution and the closeness to the exact solution.

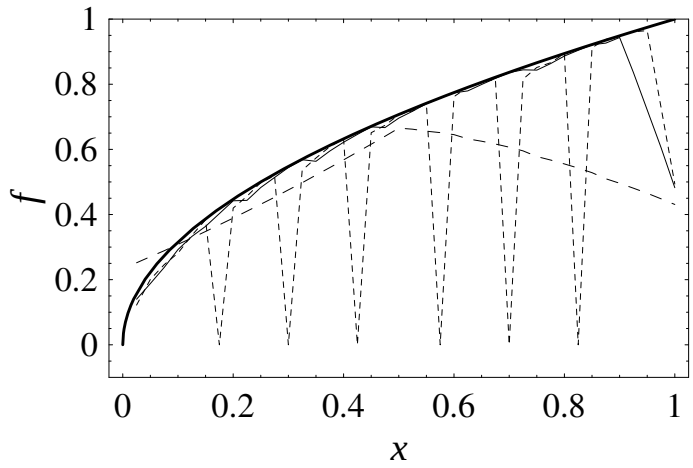


Figure 4.16. *Approximate solution versus exact solution (bold line). Dashed line:* $\lambda = 0.05$*; solid line:* $\lambda = 0.1$*, long-dashed line* $\lambda = 0.5$*. We use* $n_k = 40$*. The data number is 20.*

4.4 Rheometry: the Couette inversion

4.4.1 Introduction

Setting

The inverse problem we have to solve is the following: find the shear rate $\dot{\gamma}$ from the rotational velocity measurement Ω:

$$\Omega(M) = -\int_{\alpha_1 M}^{\alpha_2 M} \frac{1}{2}\frac{\dot{\gamma}(\tau)}{\tau}\mathrm{d}\tau, \tag{4.25}$$

where the torque is denoted by M, the shear stress by $\tau = M/(2\pi r^2)$, and the distance from the inner cylinder by r. We have also introduced $\alpha_1 = M/(2\pi h R_1^2)$ and $\alpha_2 = M/(2\pi h R_2^2)$ where R_1 and R_2 denote the inner and outer cylinder radii, whereas h denotes the flow depth.

The problem with Eq. 5.15 is that the bounds are varying. It is better to have fixed bounds, for instance, by setting:

$$g(M) = \int_0^\infty k(M,\ S)f(S)\mathrm{d}S, \tag{4.26}$$

with: $k(M,\ S) = \frac{1}{2S}H[-(S-S_2)(S_1-S)]$ and $S_i = (1+\alpha_i M)^{-1}$. In this case, however, the infinity bounds makes the numerical computations intricate. We have two strategies:

- By introducing the following change in variable: $S = (\dot{\gamma}+1)^{-1}$, it can be shown that the equation is equivalent to the following equation:

$$g(M) = \int_0^1 k(M,\ S)f(S)\mathrm{d}S, \tag{4.27}$$

  with:

$$k(M,\ S) = \frac{1}{2(1-S)}H[-(S-S_2)(S_1-S)],$$

  where H is the unit step (Heaviside) function and $S_i = (1+\alpha_i M)^{-1}$.

- We make the variables dimensionless.

We first explore the former strategy. The latter is deferred to the next section, where the problem is studied thoroughly.

Inverse and adjoint operators

Here we derive some important results characterizing the operator K:

$$(Kf)(x) = -\frac{1}{2}\int_{\alpha_1 x}^{\alpha_2 x}\frac{f(y)}{y}\mathrm{d}y$$

Let us consider f and g two functions of $\mathcal{L}(\mathbb{R})$ with specific properties to be specified. The adjoint operator can be obtained as follows:

$$< Kf,\ g >= -\frac{1}{2}\int_{\mathbb{R}}\mathrm{d}x g(x)\frac{1}{2}\int_{\alpha_1 x}^{\alpha_2 x}\frac{f(y)}{y}\mathrm{d}y$$

$$< Kf,\ g >= \frac{1}{2}\int_{\mathbb{R}} \mathrm{d}x g(x) \int_0^{\alpha_1 x} \frac{f(y)}{y}\mathrm{d}y - \frac{1}{2}\int_{\mathbb{R}} \mathrm{d}x g(x) \int_0^{\alpha_2 x} \frac{f(y)}{y}\mathrm{d}y$$

Taking the first term on the right-hand side:

$$\frac{1}{2}\int_{\mathbb{R}} \mathrm{d}x g(x) \int_0^{\alpha_1 x} \frac{f(y)}{y}\mathrm{d}y = \frac{1}{2}\left[G(x)\int_0^{\alpha_1 x} \frac{f(y)}{y}\mathrm{d}y\right]_{\mathbb{R}} - \frac{1}{2}\int_{\mathbb{R}} \mathrm{d}x G(x)\frac{f(\alpha_1 x)}{x},$$

where $G(x)$ denotes the integral of g. Here we impose the constraint:

$$G(x) \to 0 \text{ and } \int_0^{\alpha_1 x} \frac{f(y)}{y}\mathrm{d}y \to 0 \text{ when } x \to \pm\infty.$$

Making use of $z = \alpha_1$, we deduce that:

$$\frac{1}{2}\int_{\mathbb{R}} \mathrm{d}x g(x) \int_0^{\alpha_1 x} \frac{f(y)}{y}\mathrm{d}y = -\frac{1}{2}\int_{\mathbb{R}} \mathrm{d}z \frac{G(z/\alpha_1)}{z} f(z).$$

The same computation can be made with the second term, which leads to:

$$< Kf,\ g >= -\frac{1}{2}\int_{\mathbb{R}} \mathrm{d}z \frac{G(z/\alpha_1)}{z} f(z) + \frac{1}{2}\int_{\mathbb{R}} \mathrm{d}z \frac{G(z/\alpha_2)}{z} f(z) =< f,\ K^* g > .$$

The adjoint operator is then defined as:

$$\boxed{(K^* g)(x) = \frac{1}{2x}\int_{x/\alpha_1}^{x/\alpha_2} g(y)\mathrm{d}y} \tag{4.28}$$

The operator has an inverse that can be computed as follows (Coleman *et al.*, 1966). Differentiating with respect to M provides

$$\frac{\mathrm{d}\Omega(M)}{\mathrm{d}M} = \frac{1}{2}\frac{\dot\gamma(\alpha_1 M)}{M} - \frac{1}{2}\frac{\dot\gamma(\alpha_2 M)}{M},$$

or equivalently:

$$\dot\gamma(\alpha_1 M) = \dot\gamma(\alpha_2 M) + G(M), \tag{4.29}$$

where $G(M) = 2M\Omega'(M)$. We set $\beta = \alpha_2/\alpha_1$. Iteratively we have:

$$\dot\gamma(\alpha_1 M) = \dot\gamma(\alpha_1 \beta M) + G(M)$$

$$\dot\gamma(\alpha_1 \beta M) = \dot\gamma(\alpha_1 \beta^2 M) + G(\beta M)$$

$$\dot\gamma(\alpha_1 \beta^k M) = \dot\gamma(\alpha_1 \beta^{k+1} M) + G(\beta^k M)$$

By summing up to $k = N$ we obtain:

$$\dot\gamma(\alpha_1 M) = \dot\gamma(\alpha_1 \beta^{N+1} M) + \sum_{k=0}^{N} G(\beta^k M).$$

Since $\beta < 1$, $\dot\gamma(\alpha_1 \beta^{N+1} M) \to \dot\gamma(0) = 0$ and we deduce:

$$\boxed{\dot\gamma(\alpha_1 M) = \sum_{k=0}^{\infty} G(\beta^k M)} \tag{4.30}$$

The inverse of the operator K is then:

$$K^{-1}(\Omega) = \sum_{k=0}^{\infty} 2\beta^k M\Omega'(\beta^k M)$$

Its adjoint is obtained as follows:

$$< K^{-1}(f),\ g >= \sum_{k=0}^{\infty} 2\beta^k \int_{\mathbb{R}} xf'(\beta^k x)g(x)\mathrm{d}x$$

We have the relationship: $[xf(\beta^k x)g(x)]' = \beta^k xf'(\beta^k x)g(x) + f(\beta^k x)[xg]'$, thus:

$$< K^{-1}(f),\ g >= \sum_{k=0}^{\infty} 2\beta^k \left([xf(\beta^k x)g(x)]_{\mathbb{R}} - \int_{\mathbb{R}} f(\beta^k x)[xg]'\mathrm{d}x \right)$$

If f and g have finite support, then $[xf(\beta^k x)g(x)]_{\mathbb{R}} = 0$. By making the variable change $y = \beta^k x$, we deduce

$$< K^{-1}(f),\ g >= -\sum_{k=0}^{\infty} 2\int_{\mathbb{R}} f(y)[\beta^{-k} yg(y)]'\mathrm{d}x = \int_R \mathrm{d}x f(x) \left(-2\sum_{k=0}^{\infty} \beta^{-k}(xg'(x) + g(x)) \right)$$

We finally find:

$$(K^{-1})^*(g) = -2\sum_{k=0}^{\infty} \beta^{-k}(xg'(x) + g(x))$$

The last computation concerns biorthogonal vectors. Let us set: $\xi_\mu = K\psi_\mu$ where ψ_μ denotes a wavelet (μ is a set of scaling and shifting indices). We are looking for a vector η_λ such that:

$$< \eta_\lambda,\ \xi_\mu >= \delta_{\lambda,\mu}$$

Note that:

$$< \eta_\lambda,\ \xi_\mu >=< K\psi_\mu,\ \eta_\lambda >= \int_{\mathbb{R}} \mathrm{d}x(Y_\mu(\alpha_2 x) - Y_\mu(\alpha_1 x))\eta_\lambda(x),$$

where $Y_\mu = \int \psi_\mu(x)/x\mathrm{d}x$. Taking the first term on the right-hand side, we have:

$$\int_{\mathbb{R}} \mathrm{d}x Y_\mu(\alpha_2 x)\eta_\lambda(x) = [Y_\mu(\alpha_2 x)H_\lambda(x)]_{\mathbb{R}} - \int_{\mathbb{R}} \mathrm{d}x H_\lambda(x)\frac{\psi_\mu(\alpha_2 x)}{x}$$

where $H_\lambda = \int \eta_\lambda(x)\mathrm{d}x$. If we set

$$H_\lambda(x) \to 0 \text{ and } Y_\mu(x) \to 0 \text{ when } x \to \pm\infty.$$

and set $H_\lambda(\alpha_2^{-1}x) = -x\psi_\lambda(x)$, i.e. $\eta_\lambda(x) = -\alpha_2\psi_\lambda(\alpha_2 x) - \alpha_2^2 x\psi'_\lambda(\alpha_2 x)$, then we have:

$$\int_{\mathbb{R}} \mathrm{d}x Y_\mu(\alpha_2 x)\eta_\lambda(x) = \delta_{\mu,\lambda}.$$

The same can be done with the second term. So, if we define:

$$\boxed{\eta_\lambda(x) = \alpha_1\psi_\lambda(\alpha_1 x) + \alpha_1^2 x\psi'_\lambda(\alpha_1 x) - \alpha_2\psi_\lambda(\alpha_2 x) - \alpha_2^2 x\psi'_\lambda(\alpha_2 x)} \tag{4.31}$$

then the biorthogonal condition $< \eta_\lambda,\ \xi_\mu >= \delta_{\lambda,\mu}$ holds.

4.4.2 Case study 1: power fluid flow

Let us consider the case: $\dot{\gamma} = \sqrt{\tau}$. In this case, we find:

$$g(M) = (\sqrt{\alpha_2} - \sqrt{\alpha_1})\sqrt{M}$$

In the numerical simulations we took: $h = 1$, $R_1 = 1$, and $R_2 = 0.5$; 50 measurements of Ω were taken between $M_1 = 0.05$ and $M_{50} = 10$ by increments of 0.2. These values give: $\alpha_1 = 0.159$ and $\alpha_2 = 0.636$. The maximum mean shear rate (averaged over the gap) is then: $\bar{\dot{\gamma}} = \sqrt{M_{50}/(2\pi h)}(R_1 + R_2)/(R_1 R_2)^2 \approx 7.53$. So we have to explore the range 0–8 approximately.

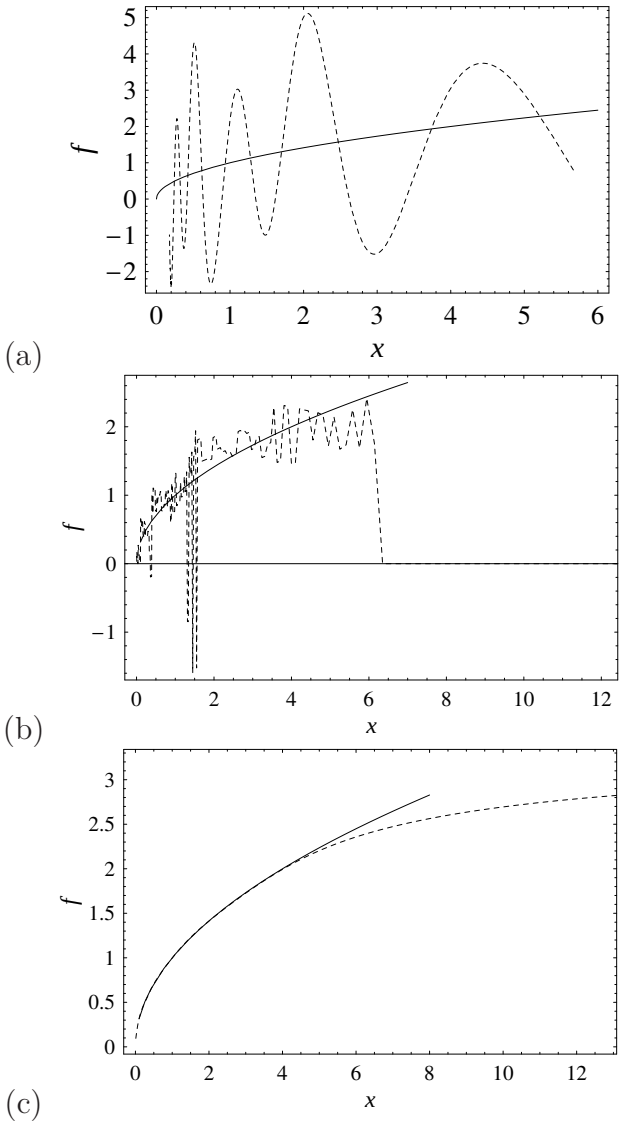


Figure 4.17. *(a) Galerkin method. (b) Direct discretization of the integral operator using the Simpson quadrature rule with n_k discretization points. (c) Tikhonov regularization of solution given in (b) with $\lambda = 10^{-8}$. In the plots, $f = \sqrt{x}$ stands for the solution (bold solid line); the dashed line represents the numerical solution.*

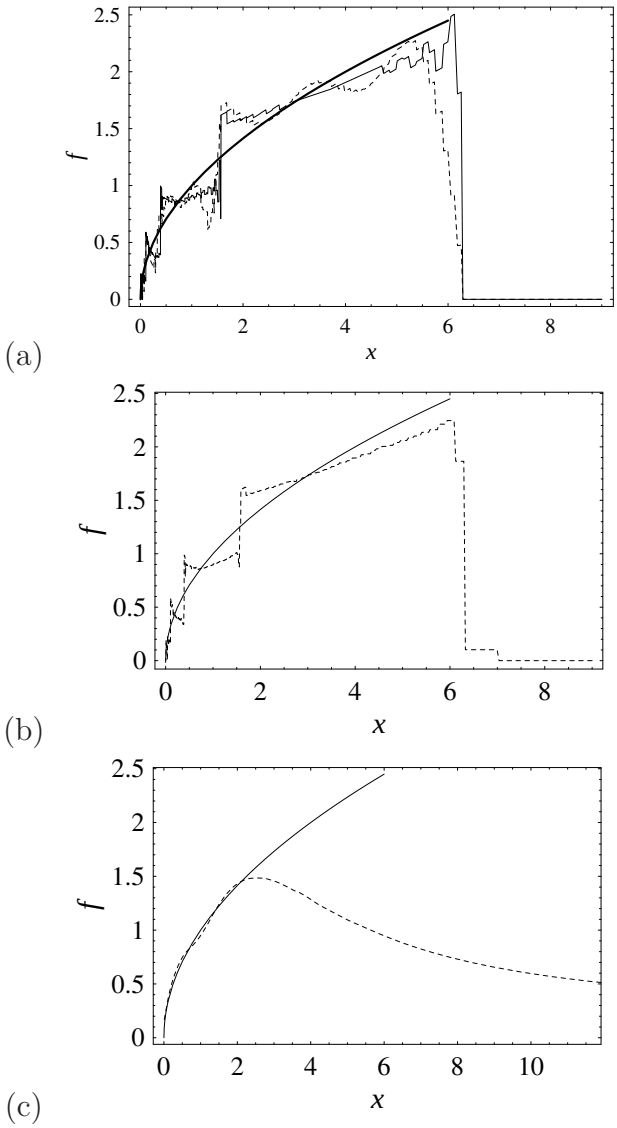


Figure 4.18. *(a) SVD method: $f = \sqrt{x}$ stands for the solution (bold solid line); the thin line represents the SVD solution, whereas the dashed line represents the truncated SVD decomposition (with a tolerance of 10^{-2} the nullity of the self-adjoint operator is 38, so we keep only the first 12 eigenfunctions). (b) WVD method. (c) Approximate inverse (mollifier method) with $n_k = 150$ discretization points and $\lambda = 0.1$; the mollifier was a Gaussian function $e_\lambda(x,\ y) = (2\pi\lambda^2)^{-1/2} \exp[-(x-y)^2/(2\lambda)^2]$.*

4.4.3 Study of a modified problem

In the Couette problem, we can seek $\Gamma(\tau) = \tau\dot{\gamma}(\tau)$ instead of $\dot{\gamma}$. In this case, the inverse problems is: finding Γ such that:

$$\Omega(M) = -\frac{1}{2}\int_{\alpha_1 M}^{\alpha_2 M} \Gamma(\tau)\mathrm{d}\tau. \tag{4.32}$$

The problem is made dimensionless by introducing:

$$\tau \to \Delta\tau t + \tau_{min},\ \Omega \to \Omega_0\omega, \text{ and } \Gamma = g\Omega_0/\Delta\tau,$$

where t, g, and ω are dimensionless variables. We also introduce $\beta = \alpha_2/\alpha_1 = (R_1/R_2)^2 < 1$, $\chi = (1-\beta)\tau_{min}/\Delta\tau$, and $\Delta\tau = \tau_{max} - \tau_{min}$, where τ_{max} and τ_{min} are the minimum and maximum measured stresses. The dimensionless bounds are:

$$S_1 = \frac{\alpha_1 M - \tau_{min}}{\Delta\tau} \text{ and } S_2 = \frac{\alpha_2 M - \tau_{min}}{\Delta\tau} = \beta S_1 - \chi.$$

Finally we have the following dimensionless equation:

$$\omega(S_1) = \frac{1}{2}\int_{\beta S_1 - \chi}^{S_1} g(\xi)\mathrm{d}\xi. \tag{4.33}$$

WVD decomposition

We denote $v_\lambda = K\Psi_\lambda$ with $\lambda = (i,\ j)$ the scaling and shifting indices. The functions $u_\mu(x) = 2\beta\Psi'_\mu(\beta x - \chi) - 2\Psi'_\mu(x)$ form a biorthogonal family: $< v_\lambda,\ u_\mu >= \delta_{\mu,\lambda}$.

5 Rheometry

Rheometry refers to a set of standard techniques that are used to experimentally determine rheological properties of materials (fluid or solid). The idea underpinning rheometry is to realize flows, where the stress and/or strain fields are known in advance, which makes it possible to deduce rheological properties from measurements of flow properties. A rheometer is usually an engine, which can exert a torque/force on a material and accurately measures its response with time (or conversely, it can impose a strain and measures the resulting torque). In this chapter, we start with a presentation of how a rheometer operates and how measurements can be used to infer the rheological properties of the material tested. Then, the experimental procedures and the typical behaviors observed are reviewed. Emphasis is also given to providing a general view on issues encountered in rheometry, either because of rheometer limitations or as a result of disturbing phenomena in the material tested.

5.1 How does a rheometer operate?

5.1.1 A long history

Originally, rheometers were based on an applied stress which was generated by a weights-and-pulleys arrangement, as shown for instance in Figure 5.1. These methods were then superseded by electrically driven motors; they were the first controlled-strain instruments. With logarithmic mechanical gear boxes, the most sophisticated rheometers (e.g. the Weissenberg Rheogoniometer) in the 1960's were able to span a shear-rate range typically from 10^{-4} to 10^{4} s^{-1}.

In the mid 1970s, a new generation of controlled-stress rheometers began to appear. The first had been developed by Deer and colleagues at the London School of Pharmacy, who used air bearings and an air-driven turbine to provide the torque.

Then around 1980, commercial versions of the new generation of electrically driven controlled-stress rheometers appeared, still based on air bearings that greatly reduced friction, but also using so-called drag-cup electrical motors that allowed controlled stresses to be more easily applied independently of rotation speed. Along with these features came new ways of measuring smaller and smaller rotation and rotation rates. The latest optical-disc technology now means that rotation rates as low as 10^{-8} rad s^{-1} (1 revolution in 20 years) can be measured! This has opened up a new range of previously unobtainable flow behaviour.

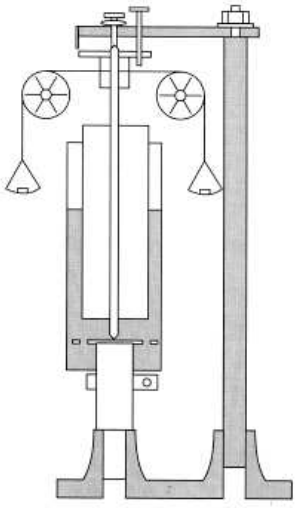


Figure 5.1. *A sketch of Searle's 1912 controlled-stress, concentric-cylinder viscometer. After (Barnes, 2000)*

5.1.2 Anatomy of a modern rheometer

At LHE, we use a CVOR 200 Bohlin rheometer. Figure 5.2 shows the heart of the rheometer when the engine hood is removed.

Figure 5.2. *Bohlin CVOR used at LHE.*

Most modern rheometers adopt the same architecture. As shown in Figure 5.3, there are several elements that are common in modern rheometers

1. Low inertia, contact-less, inductive motor assembly without any permanent magnets for accurate torque setting over a wide dynamic range and excellent transient response.

2. Precision air bearing with thrust and journal surfaces for frictionless support with high

axial and radial stiffness.

3. Inductive, contact-less and multi-layer position sensor assembly for accurate measurement of speed and absolute position.
4. Precision collett-type chuck for attachment of upper measurement system with minimal axial run-out.
5. Leadscrew assembly with bespoke pre-tensioned nut and stepper motor for accurate control of measurement sensor position with no backlash.
6. Linear bearing to support and guide the air bearing and motor assembly.
7. Integral normal force sensors placed within the instrument mechanics
8. Temperature control unit (Peltier plate design shown)
9. Universal clamping arrangement designed to accept all available temperature control options.
10. Push button acrylic key pad incorporating all key gap setting and control functions with normal force and gap separation display.
11. Integral electronics and power supply incorporating microprocessor based control and measurement functions, including torque, speed, position and temperature. Data communication is to PC via a high speed serial link.
12. Composite outer covers to provide impervious barrier and protect instrument mechanics.

Figure 5.3. *Bohlin CVOR used at LHE.*

5.1.3 Typical performance of modern lab rheometers

Modern rheometer capabilities include

- control on sample temperature;
- quite a wide range of tools (parallel-plate, cone-plane, etc.);
- wide shear-rate range ($>$ 10 orders of magnitude);
- directional (including reverse flow) and oscillatory flow;
- high accuracy and resolution;
- direct monitoring via a PC.

Here are the typical features of modern high-performance rheometer (Bohlin CVOR) :

- Torque range 0.05×10^{-6}à200×10^{-3} mN.m
- Torque resolution 1×10^{-9} Nm
- Rotational-velocity range $1 \times 10^{-7} - 600$ rad/s
- Resolution in angular position 5×10^{-8} rad
- Frequency range $1 \times 10^{-5} - 150$ Hz
- Normal force range $1 \times 10^{-3} - 20$ N

5.2 Principles of viscometry

5.2.1 Fundamentals of rheometry

Rheometry and viscometry

At the very beginning, the term *rheometry* referred to a set of standard techniques for measuring shear viscosity. Then, with the rapid increase of interest in non-Newtonian fluids, other techniques for measuring the normal stresses and the elongational viscosity were developed. Viscometry is an important offshoot of rheometry, which applies to incompressible simple fluids. When the 'simple-fluid' approximation holds, it is possible to derive the flow curve and other rheological functions (e.g., normal stress differences) from the geometrical measurements: torque, rotational velocity, and thrust.

Nowadays, rheometry is usually understood as the area encompassing any technique that involves measuring mechanical or rheological properties of a material. This includes :

- visualization techniques (such as photoelasticimetry for displaying stress distribution within a sheared material);
- nonstandard methods (such as the slump test for evaluating the yield stress of a viscoplastic material).

In most cases for applications, shear viscosity is the primary variable characterizing the behavior of a fluid. Thus in the following, we will mainly address this issue, leaving aside all the problems related to the measurement of elongational viscosity.

The basic principle of rheometry is to perform simple experiments where the flow characteristics such as the shear stress distribution and the velocity profile are known in advance and can be imposed. Under these conditions, it is possible to infer the *flow curve*, that is, the variation of the shear stress as a function of the shear rate, from measurements of flow quantities such as torque and the rotational velocity for a rotational viscometer. In fact, despite its apparent simplicity, putting this principle into practice for natural or industrial fluids raises many issues that we will discuss below. Most rheometers rely on the achievement of *viscometric* flow (Coleman *et al.*, 1966).

The simplest curvilinear flow is the simple shear flow achieved by shearing a fluid between two plates in a way similar to Newton's experiment depicted in Chap. 1. However, in practice many problems (fluid recirculation, end effect, etc.) arise, which preclude using such a shearing box to obtain accurate measurements. Another simple configuration consists of an inclined plane or a parallel-plate rheometer.

For many fluids of practical interest, viscometry is then an indispensable theory that underpins rheometrical investigation by making a clear connection between bulk measurements and rheological data. We shall see later that an incompressible simple fluid is defined as follows:

1. only isochoric motions are permitted: bulk density is constant;
2. the stress tensor $\boldsymbol{\sigma}$ is determined, to within a pressure term, by the history of the

relative deformation gradient[1] $\mathbf{F}$

$$\mathbf{s} = \boldsymbol{\sigma} + p\mathbf{1} = \mathcal{F}(\mathbf{F}(t)),$$

with $\mathbf{s}$ the extra-stress tensor, p the pressure, $\boldsymbol{\sigma}$ the stress tensor, and $\mathcal{F}$ a tensor-valued functional of $\mathbf{F}$. Time is denoted by t. This expression is called the *constitutive equation* or rheological law.

Some specific material classes can be defined (see Chap. 1):

- If the functional $\mathcal{F}$ involves the time derivative of $\mathbf{F}$ alone[2], the material is a fluid.
- If the functional $\mathcal{F}$ does not involve the time derivative of $\mathbf{F}$, the material is a solid.
- If the functional $\mathcal{F}$ is a one-to-one function, then the fluid has no memory since the stress depends on the current state of deformation alone.
- If $\mathcal{F}$ is an integral function, then the fluid behavior is characterized by memory effects: the stress state depends on the past states of deformation experienced by the material.

More complicated behaviors can be imagined, but the important point here is to recall that a wide range of behavior can be described using this formulation. For instance, if $\mathcal{F}$ involves $\mathbf{F}$ and $\mathbf{d}$, the material is said to be *visco-elastic*.

Viscometric flows

On many occasions, it is possible to create flows that induces a relative deformation gradient that is linear with time, that is, the distance between two neighboring points varies linearly with time (this distance may be zero) at any time and any point of the material. In this case, it can be shown (Coleman *et al.*, 1966) that

- There is a tensor $\mathbf{M}$, which can be interpreted as the velocity gradient and the matrix representation of which takes the form

$$\mathbf{M} = \begin{bmatrix} 0 & 0 & 0 \\ \dot{\gamma} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

for some orthogonal basis $\mathcal{B}$ and such that the relative deformation gradient $\mathbf{F}$ is $\mathbf{F}(t) = \mathbf{R}(t) \cdot (\mathbf{1} - t\mathbf{M})$, where $\mathbf{R}$ is an arbitrary orthogonal tensor, which is a function of time and satisfies $\mathbf{R}(0) = \mathbf{1}$. In the basis $\mathcal{B}$, the strain-rate and stress tensors takes the form

$$\mathbf{d} = \begin{bmatrix} 0 & \dot{\gamma} & 0 \\ \dot{\gamma} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ and } \boldsymbol{\sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & 0 \\ \sigma_{21} & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix}.$$

[1] See Chap. 1 for further information.
[2] i.e., the strain-rate tensor $\mathbf{d} = -\frac{1}{2}(\dot{\mathbf{F}} + \dot{\mathbf{F}}^T)$

- In these expressions, $\dot{\gamma}$ is the shear rate and is assumed to a control parameter. If the fluid is a simple fluid, then there is a functional $\mathcal{F}$ such that

$$\boldsymbol{\sigma} + p\mathbf{1} = \mathcal{F}(\mathbf{M}) = \mathcal{F}(\dot{\gamma}).$$

  To get rid of the pressure term (which can be determined only by solving the equations of motion, thus does not reflect any rheological property, but only isochoric constraint), we introduce

  - the shear-stress function $\tau(\dot{\gamma}) = \sigma_{21}$;
  - the first normal-stress difference $N_1 = \sigma_{11} - \sigma_{22}$;
  - the second normal-stress difference $N_2 = \sigma_{22} - \sigma_{33}$.

  These functions are called material functions since they reflect the rheological behavior of the material tested.

If a flow satisfies these conditions, it is called *viscometric*. Two subclasses are particularly important in practice:

- A *simple shear flow* is a particular case, where the shear rate is constant at any point and does vary with time. The Couette flow between two parallel, infinite, horizontal planes provides a typical example.
- More generally, curvilinear flows can be seen a generalized variant of simple-shear flows: the shear rate is permitted to vary with position, but the deformation field remains steady and two-dimensional for a certain basis.

Current geometries that allow realizing curvilinear flows are:

- simple shear flow: pressure-driven flow through parallel plates or gravity-driven flow down an inclined channel;
- vertical cylindrical tubes (Poiseuille flow): capillary rheometers;
- torsional flows: cone-and-plate and parallel-plate rheometers;
- helical flows such as flows between concentric cylinders (Couette flow): coaxial rheometers.

5.2.2 Flow down an inclined channel

To exemplify the viscometric approach, we will show how some flow properties such as the *discharge equation* (variation of the fluid discharge as a function of the flow depth) can be used to infer the constitutive equation. We consider a gravity-driven free-surface flow in a steady uniform regime down an inclined channel. The plane is tilted at an inclination θ to the horizontal. We use the Cartesian co-ordinate system of origin 0 and of basis $\mathbf{e}_x$, $\mathbf{e}_y$, $\mathbf{e}_z$ as depicted in Fig. 5.4.

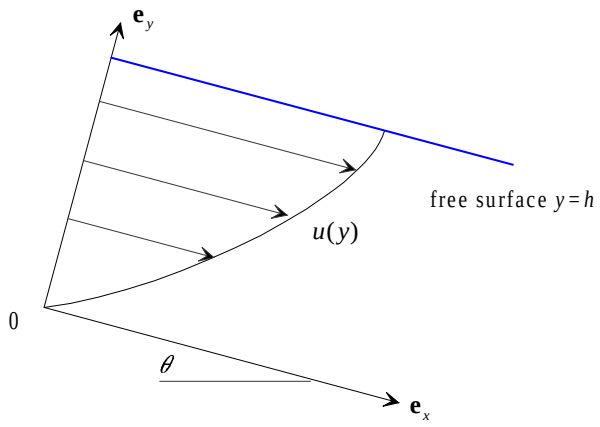


Figure 5.4. *Definition sketch for steady uniform flow*

The velocity field $\mathbf{u}$ only depends on the co-ordinate y and takes the following form: $u_x = u(y)$, $u_y = 0$, $u_z = 0$, where u is a function of y to be determined. Accordingly, the strain-rate tensor $\mathbf{d} = (\nabla\mathbf{u} + {}^t\nabla\mathbf{u})/2$ has the following components in the co-ordinate system:

$$\mathbf{d} = \frac{\dot{\gamma}}{2}\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{5.1}$$

where the shear rate $\dot{\gamma}$ is defined as a function of the co-ordinate y and implicitly of the inclination θ: $\dot{\gamma}(y) = (\partial u/\partial y)_\theta$.

The momentum balance can be written as:

$$\varrho\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = \varrho\mathbf{g} + \nabla\cdot\boldsymbol{\sigma}, \tag{5.2}$$

where ϱ and $\mathbf{g}$ respectively denote the local material density and gravitational acceleration. We assume that there is no slip at the bottom: $u(0) = 0$.

Furthermore, we assume that there is no interaction between the free surface and the ambient fluid above except the pressure exerted by the ambient fluid. Notably, we ignore surface tension effects on the free surface. Without restriction, the stress tensor can be written as the sum a pressure term p and a deviatoric term called the extra-stress tensor $\mathbf{s}$ (see also Chap. 1) (Tanner, 1988; Coleman *et al.*, 1966): $\boldsymbol{\sigma} = -p\mathbf{1} + \mathbf{s}$. For a homogeneous and isotropic simple fluid, the extra-stress tensor depends on the strain rate only: $\mathbf{s} = G(\mathbf{d})$, where G is a tensor-valued isotropic functional. In the present case, it is straightforward to show that the stress tensor must have the form

$$\boldsymbol{\sigma} = -p\mathbf{1} + \begin{bmatrix} s_{xx} & s_{xy} & 0 \\ s_{xy} & s_{yy} & 0 \\ 0 & 0 & s_{zz} \end{bmatrix}. \tag{5.3}$$

Thus, the stress tensor is fully characterized by three functions:

- the shear stress $\tau = \sigma_{xy} = s_{xy}$
- the normal stress differences: $N_1 = s_{xx} - s_{yy}$ and $N_2 = s_{yy} - s_{zz}$ called the first and second normal stress differences, respectively.

Since for steady flows acceleration vanishes and the components of $\mathbf{s}$ only depend on y, the equations of motion (5.2) reduce to

$$0 = \frac{\partial s_{xy}}{\partial y} - \frac{\partial p}{\partial x} + \varrho g \sin\theta, \tag{5.4}$$

$$0 = \frac{\partial s_{yy}}{\partial y} - \frac{\partial p}{\partial y} - \varrho g \cos\theta, \tag{5.5}$$

$$0 = \frac{\partial p}{\partial z}. \tag{5.6}$$

It follows from (5.6) that the pressure p is independent of z. Accordingly, integrating (5.5) between y and h imply that p must be written: $p(x,y) - p(x,h) = s_{yy}(y) - s_{yy}(h) + \varrho g(h - y)\cos\theta$. It is possible to express Eq. (5.4) in the following form:

$$\frac{\partial}{\partial y}\left(s_{xy} + \varrho g y \sin\theta\right) = \frac{\partial p(x,h)}{\partial x}. \tag{5.7}$$

This is possible only if both terms of this equation are equal to a function of z, which we denote $b(z)$. Moreover, Eq. (5.6) implies that $b(z)$ is actually independent of z; thus, in the following we will note: $b(z) = b$. The solutions to (5.7) are: $p\,(x,h) = bx + c$, where c is a constant, and $s_{xy}(h) - s_{xy}(y) + \varrho g(h - y)\sin\theta = b(h - y)$, which we will determine. To that end, let us consider the free surface. It is reasonable and usual to assume that the ambient fluid friction is negligible. The stress continuity at the interface implies that the ambient fluid pressure p_0 exerted on an elementary surface at $y = h$ (oriented by $\mathbf{e}_y$) must equal the stress exerted by the fluid. Henceforth, the boundary conditions at the free surface may be expressed as: $-p_0\mathbf{e}_y = \sigma\mathbf{e}_y$, which implies in turn that: $s_{xy}(h) = 0$ and $p_0 = p(x,h) - s_{yy}(h)$. Comparing these equations to former forms leads to $b = 0$ and $c = p_0 + s_{yy}(h)$. Accordingly, we obtain for the shear and normal stress distributions

$$\tau = \varrho g(h - y)\sin\theta, \tag{5.8}$$

$$\sigma_{yy} = s_{yy} - (p - p_0) = -\varrho g(h - y)\cos\theta. \tag{5.9}$$

The shear and normal stress profiles are determined regardless of the form of the constitutive equation. For simple fluids, the shear stress is a one-to-one function of the shear rate: $\tau = f(\dot\gamma)$. Using the shear stress distribution (5.8) and the inverse function f^{-1}, we find: $\dot\gamma = f^{-1}(\tau)$. A double integration leads to the flow rate (per unit width):

$$q = \int_0^h \int_0^y f^{-1}(\tau(\xi))d\xi = \int_0^h u(y)\mathrm{d}y. \tag{5.10}$$

An integration by parts leads to:

$$q(h,\ \theta) = \left[(y - h)u(y)\right]_0^h + \int_0^h (h - y)\left(\frac{\partial u}{\partial y}\right)_\theta \mathrm{d}y.$$

In this equation, the first expression of the right-hand term is hu_g if the slip condition at the bottom is relaxed. Making use of the shear stress equation leads to

$$q(h,\theta) = hu_g + \int_0^h (h - y) f\left(\varrho g \sin\theta(h - y)\right)\mathrm{d}y$$

By making the variable change: $\zeta = h - y$, we also obtain

$$q(h,\theta) = \int_0^h \zeta f\left(\varrho g \sin\theta \upsilon\right) \mathrm{d}\zeta + h u_g.$$

Thus the partial derivative of q with respect to h (at a given channel slope θ) is

$$\left(\frac{\partial q}{\partial h}\right)_\theta = h f(\varrho g h \sin\theta) + u_g + h\left(\frac{\partial u_g}{\partial h}\right)_\theta,$$

or equivalently

$$f(\tau_p) = \frac{1}{h}\left(\frac{\partial q}{\partial h}\right)_\theta - \frac{u_g}{h} - \left(\frac{\partial u_g}{\partial h}\right)_\theta.$$

where $\tau_p = \varrho g \sin\theta$ is the bottom shear shear. In the case (often encountered) of no-slip, this expression reduces to

$$\dot{\gamma} = f^{-1}(\tau(h)) = \frac{1}{h}\left(\frac{\partial q}{\partial h}\right)_\theta. \tag{5.11}$$

This relation allows us to directly use a channel as a rheometer.

The other normal components of the stress tensor cannot be easily measured. The curvature of the free surface of a channelled flow may give some indication of the first normal stress difference. Let us imagine the case where it is not equal to zero. Substituting the normal component s_{yy} by $s_{yy} = s_{xx} - N_1$ in (5.5), after integration we find:

$$s_{xx} = p + \varrho g y \cos\theta + N_1 + c, \tag{5.12}$$

where c is a constant. Imagine that a flow section is isolated from the rest of the flow and the adjacent parts are removed. In order to hold the free surface flat (it will be given by the equation $y = h$, $\forall z$), the normal component σ_{xx} must vary and balance the variations of N_1 due to the presence of the sidewalls (for a given depth, the shear rate is higher in the vicinity of the wall than in the center). But at the free surface, the boundary condition forces the normal stress σ_{xx} to vanish and the free surface to bulge out. To first order, the free surface equation is:

$$-\varrho g y \cos\theta = N_1 + c. \tag{5.13}$$

If the first normal stress difference vanishes, the boundary condition $-p_0 \mathbf{e}_y = \sigma \mathbf{e}_y$ is automatically satisfied and the free surface is flat. In the case where the first normal stress difference does not depend on the shear rate, there is no curvature of the shear free surface. The observation of the free surface may be seen as a practical test to examine the existence and sign of the first normal stress difference and to quantify it by measuring both the velocity profile at the free surface and the free-surface equation.

5.2.3 Standard geometries

Computation of the shear-stress function and normal stress differences is very similar for other types of viscometers. Figure 5.5 reports the corresponding functions for the most common viscometers. All these techniques are robust and provide accurate measurements for classic fluids, with uncertainty usually less than 2%.

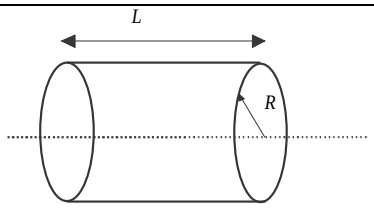
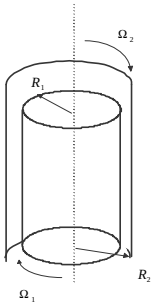
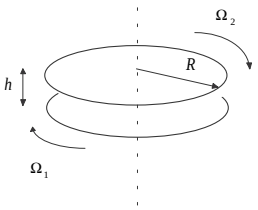
Rheometer type	Sketch	Viscometric function
Capillary tube (Poiseuille flow)		$\dot{\gamma} = \lambda\left(\frac{\Delta p_g}{L}\frac{R}{2}\right) = \frac{1}{\pi\alpha^2 R^3}\frac{\partial\left(q\alpha^3\right)}{\partial\alpha}$ $\alpha = \frac{\Delta p_g}{L}$ (pressure variation per unit length) q: flow rate, Δp_g: applied pressure gradient
Concentric cylinder (Couette flow)		$\Delta\Omega = \Omega_2 - \Omega_1 = \frac{1}{2}\int_{C/(2\pi R_2^2)}^{C/(2\pi R_1^2)} \dot{\gamma}(\tau) d\ln\tau$ $\tau = \frac{C}{2\pi R_1^2}$ R_1: inner radius; R_2: outer radius C: torque (per unit height)
Parallel-plate		$\dot{\gamma} \approx R\frac{\Delta\Omega}{h}$, $C = \frac{M}{2\pi R^3}$, $\Delta\Omega = \Omega_2 - \Omega_1$ $\tau = C\left(3 + \frac{\partial\ln C}{\partial\ln\dot{\gamma}}\right)$ M: measured torque
Inclined plane	See Fig. 1	$\tau = \rho g h \sin\theta$ $\dot{\gamma} = \frac{1}{h}\left(\frac{\partial q}{\partial h}\right)$

Figure 5.5. *Chief geometries used in rheometry*

5.3 Inverse problems in rheometry

5.3.1 A typical example: the Couette problem

A longstanding problem in rheometry is the so-called Couette inverse problem, in which one tries to derive the flow curve $\tau(\dot{\gamma})$ from the torque measurements $M(\omega)$ in a coaxial cylinder (Couette) rheometer, where τ is the shear stress, $\dot{\gamma}$ denotes the shear rate, ω is the rotational velocity of the inner cylinder, and M represents the torque per unit height (Coleman *et al.*, 1966). The shear stress τ exerted on the inner cylinder of radius R_1 can be directly related to the measured torque M by $\tau = \alpha_1 M$, with $\alpha_1 = 1/(2\pi R_1^2)$, independently of the form of the constitutive equation. The shear rate is related to the rotational velocity ω by

$$\omega = \int_{R_1}^{R_2} \frac{\dot{\gamma}(r)}{r} dr, \tag{5.14}$$

where R_2 denotes the outer-cylinder radius and it is assumed that (i) the rotational velocity of the outer cylinder is zero and (ii) there is no slip between the inner cylinder and the sheared material at $r = R_1$. In order to recover the flow curve from measurements of the rotational velocity $\omega(M)$, one must be able to

(i) relate the function $\dot{\gamma}(r)$ to $\tau(r)$,

(ii) find out a means of inverting the integral relationship (5.14),

(iii) estimate the continuous function $\dot{\gamma}(\tau)$ from a set of discrete values $(\omega_i,\ M_i)$.

For a broad class of fluids (simple fluids), the first step is systematically achieved since there is a one-to-one relation between the shear stress and the shear rate for steady viscometric flows: $\dot{\gamma} = \dot{\gamma}(\tau)$. Moreover, the momentum equations imply that the shear stress distribution across the gap is given by $S(r) = M/(2\pi r^2) = \tau(R_1/r)^2$, where r denotes the distance from the vertical rotation axis of the cylinders. Under these conditions, which are not too stringent, it is possible to make the variable change $r = R_1\sqrt{\tau/S}$ in the integral above; we then derive the well-known equation (Krieger & Elrod, 1953; Coleman *et al.*, 1966)

$$\omega(\tau) = \frac{1}{2}\int_{\beta\tau}^{\tau} \frac{\dot{\gamma}(S)}{S} dS, \tag{5.15}$$

where $\beta = (R_1/R_2)^2$. The next step is to recover $\dot{\gamma}$ from $\omega(\tau)$.

5.3.2 Earlier attempts at solving the Couette problem

Scientific statement and mathematical strategies

In the Couette inverse problem, Eq. (5.15) can be represented in the generic form: $\omega(\tau) = (K\dot{\gamma})(\tau)$, where K is the integral operator

$$(Kf)(z) = \int_{\beta z}^{z} \frac{f(x)}{x} dx, \tag{5.16}$$

with β a constant parameter ($\beta < 1$). A considerable body of literature has been published over the last three decades on ill-posed inverse problems in this form (Bertero *et al.*, 1985,

1988; O'Sullivan, 1986; Tenorio, 2001). Schematically, we can split the various methods for solving Couette-like problems into three main categories[3].

- *Least-square approach*: instead of solving $\omega = K\dot{\gamma}$, an attempt is made to minimize the residual $||\omega - K\dot{\gamma}||$, usually with an additional constraint on the norm of $||f||$ or its derivative(s), to control the smoothness of the solution. Tikhonov's regularization method used by Yeow *et al.* (2000) and Landweber's iterative procedure used by Tanner & Williams (1970) come within this category. The advantages of this method are its robustness against computation inaccuracies and measurement errors, its versatility, its fast convergence when the function to be recovered behaves reasonably well, and the relative facility of its implementation. The drawbacks are that it relies on an arbitrary selection of the regularization operator (even though specific procedures have been established) and its limited capacity to retrieve irregular functions.

- *Projection approach*: the idea here is to discretize the problem by projecting the function over a finite space spanned by a family of functions enjoying specific properties (such as orthogonality) u_i. Equation (5.15) is then replaced by the finite set of equations $\langle K\dot{\gamma},\ u_i\rangle = \langle \omega,\ u_i\rangle$ for $1 \leq i \leq p$, where $\langle f,\ g\rangle = \int_{\mathbb{R}} f(x)g(x)dx$ denotes the inner product in the function space (Dicken & Maass, 1996; Louis *et al.*, 1997; Rieder, 1997). Galerkin's method, used by Macsporran (1989) with spline functions, provides a typical example for Couette rheometry. Irregular functions can be recovered by these methods provided appropriate projection functions are chosen in advance.

- *Adjoint operator approach*: for many reasons, it is usually either not possible or not advantageous to compute the inverse operator K^{-1}. In some cases, however, it is possible to provide a weak inverse formulation, in which the function $\dot{\gamma}$ is expressed as

$$\dot{\gamma} = \sum_{i \in J} \langle K\dot{\gamma},\ u_i\rangle \Psi_i,$$

  where the summation is made over a set J, Ψ_i is an orthonormal basis of functions, and u_i denotes a family of function solutions of the adjoint problem $K^* u_i = \Psi_i$, where K^* is the adjoint operator of K (Golberg, 1979). Typical examples include *singular-value decomposition* (Bertero *et al.*, 1985, 1988), a generalized formulation based on *reconstruction kernels* (Louis, 1999), *wavelet-vaguelette decomposition* (Donoho, 1995), and *vaguelette-wavelet decomposition* (Abramovich & Silverman, 1998). The solution to the inverse problem is found by replacing $K\dot{\gamma}$ with ω in the equation above and filtering or smoothing the inner products $\langle K\dot{\gamma},\ u_i\rangle$ and/or truncating the sum.

Mooney's and Krieger's approximation

Although the Couette problem admits an analytical theoretical solution in the form of an infinite series (Coleman *et al.*, 1966), deriving the shear rate remains a difficult task in practice because the derivation enters the class of ill-posed problems (Friedrich *et al.*, 1996). In

[3] This partitioning is a bit arbitrary because there are interconnections between the three categories [e.g., Tikhonov's regularization can be viewed as a special case of singular-value decomposition (Bertero *et al.*, 1988)]. This is, however, sufficient in the present paper to outline the main approaches used so far and to situate the previous attempts at solving the Couette problem. Alternative methods, e.g., stochastic methods (Gamboa & Gassiat, 1997; Mosegaard & Sambridge, 2002), are also possible, but have never been used in rheometry as far as we know.

rheometry, the first attempt at solving Eq. (5.15) can be attributed to Mooney (1931), Krieger & Maron (1952), and Krieger & Elrod (1953). When β is close to unity, it is possible to directly approximate the integral to the first order by

$$\omega(\tau) = \frac{1-\beta}{2}\dot{\gamma}(\tau) + o(\beta\dot{\gamma}).$$

When β moves away from unity, further terms are needed in the expansion of the integral into a β series. One of the most common approximations is attributed to Krieger who proposed for Newtonian and power-law fluids (Yang & Krieger, 1978; Krieger, 1968):

$$\dot{\gamma} = \frac{2\Omega(1+\alpha)}{1-\beta^f} f, \tag{5.17}$$

with

$$f = \frac{\mathrm{d}\ln\Omega}{\mathrm{d}\ln C},\ \alpha = \frac{f'}{f^2}\chi_1(-f\log\beta), \text{ and } \chi_1(x) = \frac{x}{2}(xe^x - 2e^x + x + 2)(e^x-1)^{-2}.$$

However, this method can give poor results with yield stress fluids, especially if it is partially sheared within the gap. In this case, Nguyen & Boger (1992) have proposed using

$$\dot{\gamma} = 2\Omega\frac{\mathrm{d}\ln\Omega}{\mathrm{d}\ln C}.$$

A few rheologists used an alternative consisting of an expansion into a power series of (5.15). They obtained:

$$\dot{\gamma} = 2\Omega\sum_{n=0}^{\infty} f\left(\beta^n C/(2\pi R_1^2)\right).$$

Although refined to achieve higher accuracy (Yang & Krieger, 1978), Krieger's approach was unable to provide reliable results for viscoplastic flows (Darby, 1985; Nguyen & Boger, 1992) or for data contaminated by noise (Borgia & Spera, 1990).

Tikhonov's regularization technique

Alternative methods have been proposed: Tanner & Williams (1970) developed an iterative procedure, whereas Macsporran (1989), Yeow *et al.* (2000), and Leong & Yeow (2003) used a regularized least-square approach, which involves discretizing the integral term and regularizing it.

These methods are very efficient for a wide range of well-behaved rheological equations. However, when the rheological behavior exhibits singularities such as a yield stress or a rapid shear-thickening, the regularization procedure can lead to unrealistic results by smoothing out the singularities or to complicated trial-and-error loops. For instance, when testing Tikhonov's method with viscoplastic flows, Yeow *et al.* (2000) had to evaluate the yield stress iteratively, which may involve a large number of computations and slow convergence. This undesired behavior is to a large extent the result of attempting to evaluate a continuous function ($\dot{\gamma}(\tau)$) from a finite set of discrete values representing measurements of bulk quantities. This task is more delicate than believed, especially when data are noisy. For a well-behaved rheological equation, imposing a certain degree of smoothness in the regularization procedures does not entail many problems. On the contrary, for complex rheological responses, it becomes increasingly difficult to discern genuine rheological properties, noise effects, and discretization errors.

5.3.3 The wavelet-vaguelette decomposition

We will begin by exposing the principle in a very simple manner. A more rigorous mathematical derivation follows in the Appendix. Let us assume that we can approximate any shear rate function $\dot{\gamma}(\tau)$ with a finite series of terms

$$\dot{\gamma}(\tau) \approx \sum_k a_k \Psi_k(\tau),$$

where Ψ_k denotes the k^{th} member of a family of orthogonal functions, i.e., $\int \Psi_k(\tau)\Psi_i(\tau)d\tau = \delta_{ik}$; making use of this property, we could compute the coefficients a_k as $a_k = \int \dot{\gamma}(\tau)\Psi_i(\tau)d\tau$ if the function $\dot{\gamma}(\tau)$ were known.

Using the linearity of the integral operator K, we have

$$\omega(\tau) = (K\dot{\gamma})(\tau) \approx \sum_k a_k \, (K\Psi_k)(\tau).$$

Note that the function $\omega(\tau)$ shares the same coefficients a_k as the shear-rate function, implying that if we were able to expand $\omega(\tau)$ into a $(K\Psi_k)$ series, we could determine the coefficients a_k, then find an approximation of $\dot{\gamma}(\tau)$.

Unfortunately, the functions $(K\Psi_k)(\tau)$ are not orthogonal, making it difficult to numerically compute a_k. Specific procedures such as the Schmidt orthogonalization procedure could be used to derive an orthogonal family of functions from $(K\Psi_k)(\tau)$, but here this involves overly complicated computations. We will envisage another technique based on dual bases. A dual basis of the function basis Ψ_k is a set of functions u_i such that $\int u_i(\tau)(K\Psi_k)(\tau)d\tau = \delta_{ik}$, implying that $a_k = \int \omega(\tau)u_k(\tau)d\tau$. Therefore the crux of the issue lies in the derivation of the dual basis u_k. In the following, we will show that the functions u_k can be built from the functions Ψ_i.

5.3.4 Practical example

Baudez *et al.* (2004) investigated the rheological properties of a polymeric suspension (commercial hair gel made of Carbopol) using a stress-controlled Paar Physica MC1+ rheometer equipped with a Couette geometry ($R_1 = 1.25$ cm and $\beta = 0.26$). In addition they carried out velocity-profile measurements in a similar geometry ($R_1 = 4$ cm and $\beta = 0.44$) using magnetic resonance imaging (MRI) techniques. Further rheometrical tests were also done with a Bohlin CVOR200 rheometer ($R_1 = 0.0125$ cm and $\beta = 0.06$). Carbopol suspensions usually exhibit a viscoplastic behavior (Roberts & Barnes, 2001). MRI techniques made it possible to obtain an accurate estimation of the flow curve and then to compare the different methods.

The data obtained by Baudez *et al.* (2004) are reported in a log-linear plot in Fig. 5.6. They were slightly noisy and a specific procedure was used to denoise and interpolate the raw data. Different nonparametric regression techniques can be used for this purpose: kernel estimator (Hart, 1999), spline smoothing (Wahba, 1990), Fourier series estimator, wavelet regression and shrinkage (Donoho & Johnstone, 1995; Kovac, 1998; Cai, 1999, 2002), Bayesian inference (Werman & Keren, 2001), etc. There is not a universal method because, depending on the noise level, the number of data, and the properties of the function to be recovered, the performance of each method in terms of denoising efficiency can vary significantly. Here, because of the small size of the data samples, the optimized Gasser-Müller kernel method

(included in the Mathematica package) was used to denoise and interpolate the data [for details in the implementation, see (Hart, 1999)]. The resulting interpolating curves are plotted in Fig. 5.6.

Figure 5.7 shows the flow curves deduced by the Tikhonov regularization method (dashed line) and the wavelet-vaguelette decomposition method (solid line) for the polymeric gel. For the Tikhonov method, we used the method described in (Yeow *et al.*, 2000) with $n_k = 400$ discretization points and a smoothing parameter $\lambda = 2 \times 10^{-6}$ and 5×10^{-6}. For the WVD method, Daubechies D8 wavelet and the functional formulation were used.

For the polymeric gel, it was possible to independently obtain a reference flow curve by using the velocity profile determined by Baudez *et al.* (2004) using MRI techniques. Indeed, in a Couette geometry, the shear stress distribution across the gap is imposed: $\tau(r) = M/(2\pi r^2)$; the shear rate can be computed by differentiating the velocity profile $v(r)$: $\dot{\gamma}(r) = -r\partial(v/r)/\partial r$. Reporting a parametric plot $(\dot{\gamma}(r), \tau(r))$ as a function of the radial distance r makes it possible to have a clearer idea on the flow curve for the material tested. The dots in Fig. 5.7 represent the flow curve determined in this way.

For the polymeric gel [see Fig. 5.7], the three methods compare well over a shear-rate range covering approximately two orders of magnitude ($5 \times 10^{-2} \leq \dot{\gamma} \leq 20$ s^{-1}), whereas differences can be observed at low and high shear rates. Because of the smoothing constraint imposed on the flow curve in the Tikhonov method, the shear stress drops quickly at low shear rates, leading to an underestimation of the yield stress (estimated at 41 Pa using independent tests). Similarly, at large shear rates, the slight convexity of the flow curve (in a log-linear representation) leads to an undue increase in the shear stress. Because of the absence of regularization constraint in the WVD method, the corresponding flow curve comes closer to the experimental flow curve inferred from MRI Measurements. We can, however, notice the bump for shear rates in the range $10^{-3} - 5 \times 10^{-2}$, which seems not natural. This is probably an artifact caused by the interpolating curve [see Fig. 5.6] since a similar bump is also observable. Additional rotational-velocity data are required to improve accuracy in the low-shear-rate limit.

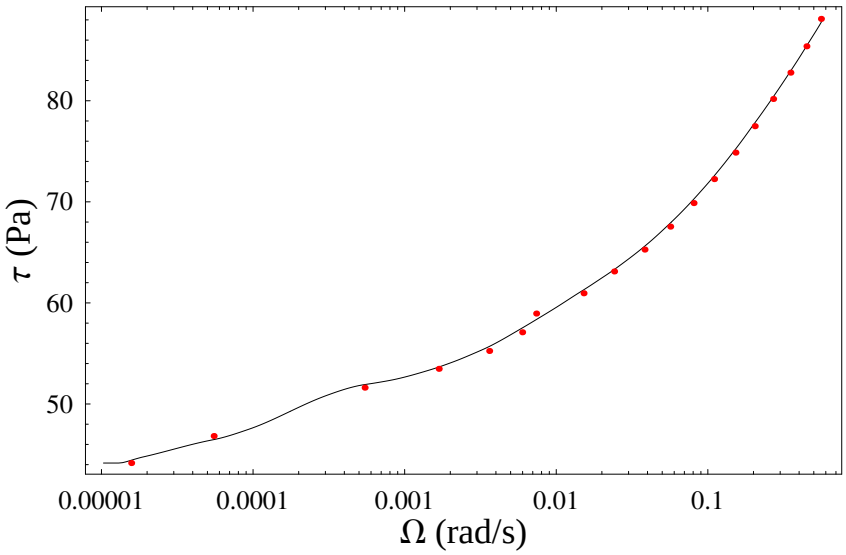


Figure 5.6. *Raw data obtained by Baudez et al. (2004) for a polymeric gel (Carbopol). Dots correspond to data while the solid lines represent the curve interpolating the data obtained using the Gasser-Müller kernel method (bandwidth parameter taken at 0.1).*

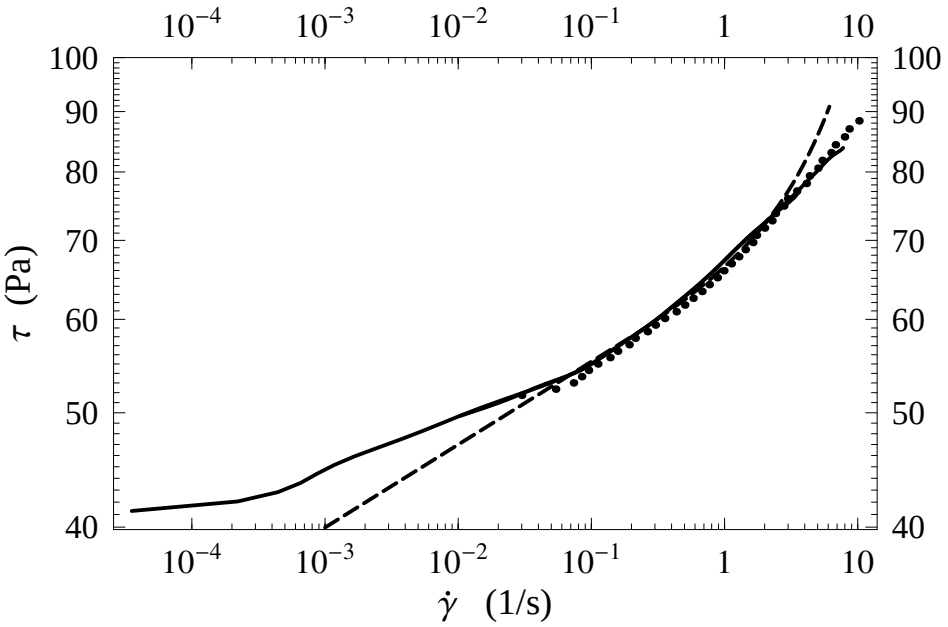


Figure 5.7. *Flow curve for the polymeric gel. The dots represent the flow curve obtained by differentiating the MRI velocity profile. The dashed lines represent the flow curves obtained using the Tikhonov method [smoothing parameter $\lambda = 2 \times 10^{-6}$]. The solid lines represent the flow curve determined using the WVD method.*

5.4 Rheometers and rheometrical procedures

In shear flows, the stress distribution is fully characterized when we know the shear stress $\tau = \sigma_{xy}$ and the normal stresses σ_{xx}, σ_{yy}, and σ_{zz}. A familiar and equivalent characterization is based on the use of the normal stress differences

$$N_1(\dot{\gamma}) = \sigma_{xx} - \sigma_{yy} \text{ and } N_2(\dot{\gamma}) = \sigma_{yy} - \sigma_{zz},$$

which makes it possible to get rid of the pressure term (recall that $\sigma_{xx} = -p + s_{xx}$, where s_{xx} is the extra stress).

The bulk viscosity is

$$\mu(\dot{\gamma}) = \frac{\tau}{\dot{\gamma}}.$$

Similarly, we introduce the normal-stress coefficients

$$\Psi_1 = \frac{N_1}{\dot{\gamma}^2} \text{ and } \Psi_2 = \frac{N_2}{\dot{\gamma}^2}.$$

$\rightsquigarrow$ We want to determine $\mu = f(\dot{\gamma})$ (*viscosity curve*) or equivalently $\tau = f(\dot{\gamma})$ (*flow curve*). In addition, normal stress differences may be of importance. Most often only $N_1(\dot{\gamma})$ can be measured (e;g., with a cone-plane rheometer).

5.4.1 How to determine the flow curve?

Manual or automatic procedure?

Most rheometers are now controlled by a computer, which also provides additional softwares for computing the flow curve automatically. When the experimentalist has a sufficiently good knowledge of the rheological properties of the material tested and the viscometric geometry used in the testing is standard, these softwares are very helpful.

For complex materials or for non-standard geometries, it is usually better to directly extract the measurement data (torque, rotational velocity) and use specific methods to determine the flow curves, which makes it possible:

- to use a specific experimental protocole in data acquisition or processing;
- to modify the data to take disturbing phenomena into account;
- to obtain more accurate solutions (e.g., in the inverse problem for wide-gap rheometers, by controlling the smoothness of the solution sought).

For complex fluids (the general case for natural fluids studied in geophysics and in industry), rheometry is far from being an ensemble of simple and ready-for-use techniques. On the contrary, investigating the rheological properties of a material generally requires many trials using different rheometers and procedures. In some cases, visualization techniques (such as nuclear magnetic resonance imagery, transparent interstitial fluid and tools, birefringence techniques) may be helpful to monitor microstructure changes.

How to measure the flow curve?

In most modern rheometers, the standard technique involves imposing a step-like *ramp*, i.e., a succession of stress steps (respectively, strain steps), and measuring the resulting deformation (respectively, stress). It is difficult to prescribe the duration of each step in advance because it basically depends on how quickly the material reaches its steady state.

Plotting the data

The data obtained by using a rheometer usually cover a wide range of shear rate, typically 3 or more orders of magnitude. For this reason, it is usually recommended to plot the data on a logarithmic basis. When the data are plotted, different things can be done:

- simple mathematical expressions can be fitted to the data, e.g., a power-law relation: $\tau = K\dot{\gamma}^n$;
- a yield stress can be evaluated by extrapolating the experimental curve to $\dot{\gamma} = 0$ and give an apparent intercept on the τ-axis that can be interpreted as a yield stress.

Note that the flow curve $\tau = f(\dot{\gamma})$ can be plotted as:

- shear stress as a function of shear rate
- bulk viscosity $\mu = \tau/\dot{\gamma}$ as a function of shear rate

Recall that the units used in viscometry are the following

- strain in %
- stress in Pascal [Pa] 1 Pa = kg$\times$m$^2\times$s^{-2}
- shear rate in 1/s (and not Hertz)
- dynamical viscosity μ in [Pa.s] 1 Pa.s = kg$\times$m$^2\times$s^{-1} (avoid Poise, centiPoise, etc.)
- density ϱ in [kg/m^3]
- dynamical viscosity $\nu = \mu/\varrho$ in [m^2/s] (avoid Stokes, centiStokes [cS], etc.)

5.4.2 Stress/strain step

Basic experiments and probably the simplest we can think of is (see Figure 5.8)

- to suddenly exert a stress on a material at rest over a sufficiently long time, then to measure the strain output after flow cessation: *recovery* test.
- to suddenly exert a stress on a material at rest over a sufficiently long time, then to measure the strain output after flow inception: *creep* test.
- to suddenly impose a steady shear flow, then to monitor the stress variation with time to determine how the shear stress reaches its steady value: *stress growth* test.

- to suddenly impose a steady shear flow, keep it constant over a given time interval, then cease the flow and monitor the stress variation with time after flow cessation (fluid at rest): *stress relaxation* test.
- to realize a steady shear flow over a given time interval, then remove the shear stress and monitor the strain variation with time: *constrained recoil* test. A viscoelastic material recoils because of elasticity, whereas a Newtonian fluid stops immediately.

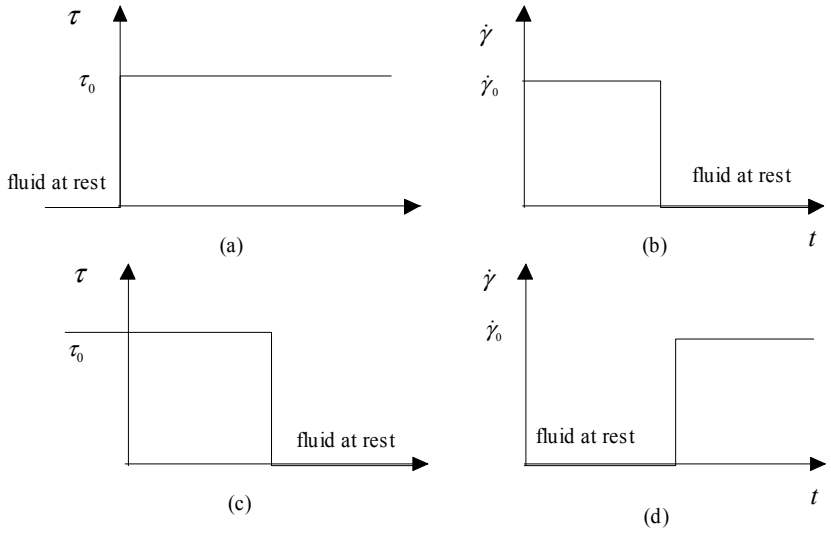


Figure 5.8. *Basic tests: (a) Creep. (b) Stress relaxation. (c) Recovery. (d) Stress growth.*

Plotting the time variation of the output signal makes it possible to exhibit some fundamental properties such as viscoelasticity or thixotropy (see below). Figure 5.9 shows the typical response for a Newtonian fluid (solid line) and a non-Newtonian fluid (dashed line). For both materials, the shear stress tends towards a limiting value, which means that the material has reached a new equilibrium (steady state). However, how the stress reaches this limiting value differs depending on the material:

- for a Newtonian fluid, the stress reaches its steady state quasi instantaneously;
- for a non-Newtonian fluids, it is common to observe an overshoot, then a decrease towards the steady value.

For a non-Newtonian fluid, the *overshoot* can be understood as follows: when the material has a structure on the microscopic scale (e.g., polymers connection, particle network, etc.), deforming the material implies that the structure must be re-organized, e.g., by breaking contacts between particles in close contact for a suspension: more energy must be provided to the system for it to reach a new equilibrium. For a thixotropic material, the time needed to reach this equilibrium depends on the previous states (intensity of the shear rate, duration of the resting procedure).

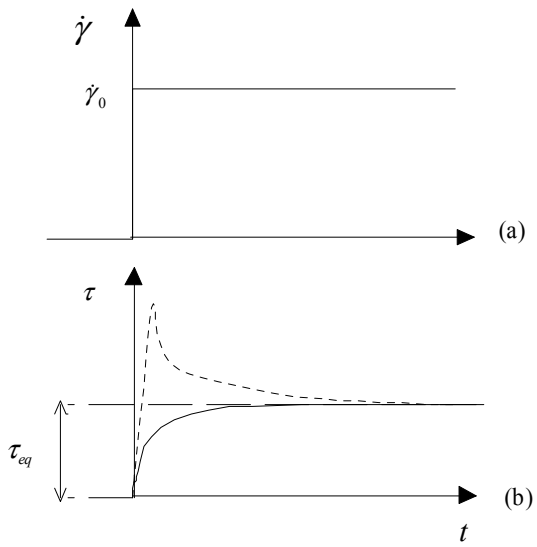


Figure 5.9. *Stress growth: shear rate imposed at $t = 0$ and stress response measured upon flow inception. (a) Input: constant shear rate imposed at $t = 0$. (b) Output: time variation of τ monitored upon flow inception. How the shear stress reaches the steady-state value τ_{eq} depends on the rheological properties: the typical response of a Newtonian fluid (solid line) and a visco-elastic material (dashed line) is depicted.*

5.5 Typical rheological behaviors

5.5.1 Outlining a flow curve

The *flow curve* is the relation between the shear rate $\dot{\gamma}$ and shear stress τ established from experimental measurements taken in a viscometric flow (i.e., meaning that a simple shear flow was realized by appropriate means and that we are able to derive the $\dot{\gamma} - \tau$ curve). On many occasions, the flow curve is represented in the form

$$\mu = \frac{\tau}{\dot{\gamma}} = f(\dot{\gamma}),$$

where f is a function that we want to characterize.

5.5.2 Shear-thinning/thickening

Many fluids exhibit the same kind of behaviour (see Fig. 5.10):

- at low shear rates, the viscosity is constant and we say that the viscosity lies in the Newtonian plateau;
- for increasing shear rate, the viscosity decreases (i.e., $f'(\dot{\gamma}) < 0$), the behavior is said to be *shear-thinning*. On a log-log plot, this trend is represented by a straight line when the fluid behavior can be described with a *power-law model* $f(\dot{\gamma}) = K\dot{\gamma}^n$, with $n < 1$ the *power-law index* and K the *consistency*;
- at high shear rates, the viscosity curve may start flattening out and reach another plateau.

On rare occasions, the viscosity is seen to increase with shear rates and in that case, the behavior is said to be *shear-thickening*[4].

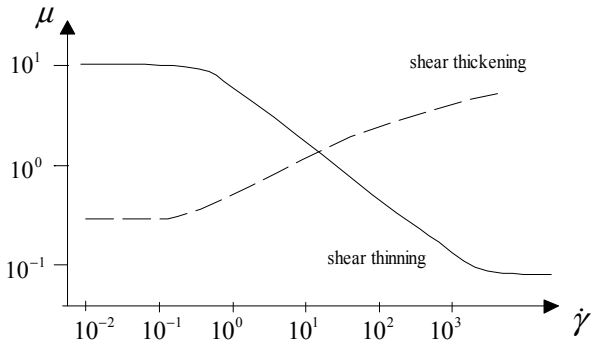


Figure 5.10. *Sketch of a shear-thinning/thikening flow.*

The Cross model is well appropriate for fitting shear-thinning fluids with two plateaux

$$\frac{\mu - \mu_\infty}{\mu_0 - \mu_\infty} = \frac{1}{1 + K\dot{\gamma}^n},$$

with μ_0 and μ_∞ the viscosities at low and high shear rates. Other similar models are the Carreau model

$$\frac{\mu - \mu_\infty}{\mu_0 - \mu_\infty} = \frac{1}{(1 + K\dot{\gamma}^2)^m},$$

and the Sisko model (a simplified version of the Cross model when $\mu_0 \gg \mu_\infty$)

$$\mu = \mu_\infty + \frac{\mu_0}{K\dot{\gamma}^n} \text{ or } \tau = \mu_\infty + \frac{\mu_0}{K}\dot{\gamma}^{1-n}.$$

5.5.3 Yield stress

Definition

For some fluids, the flow curve when plotted in the $\dot{\gamma} - \tau$ plane exhibits a *yield stress*: when the rate is decreased towards zero, the shear stress tends towards a constant value. Although extrapolating to zero is not possible when working in log-log plot, a common practice is to consider the limiting stress as a yield stress, i.e. the stress threshold below which there is no motion ($\dot{\gamma} = 0$). As shown in Fig. 5.11, the idea is to extrapolate the experimental trend towards $\dot{\gamma} = 0$. There are many problems around the interpretation of the yield stress determined in this way [e.g. see papers in the Journal of Rheology echoing the debate around the relevance of the yield stress (Harnett & Hu, 1989; Astarita, 1990; Evans, 1992; de Kee & Chan Man Fong, 1993; Spaans & Williams, 1995; Barnes, 1999)]. Care must taken in defining the yield stress as the intercept of the extrapolated flow curve with the τ-axis, especially when the experimentalist uses a rheometer with a limited range of shear rates (typically in excess of 0.1 s^{-1}). Additional tests are recommended (typically creeping test with a controlled-stress rheometer).

[4]Not to confuse with dilatancy since in the past, the two expressions were used with same meaning.

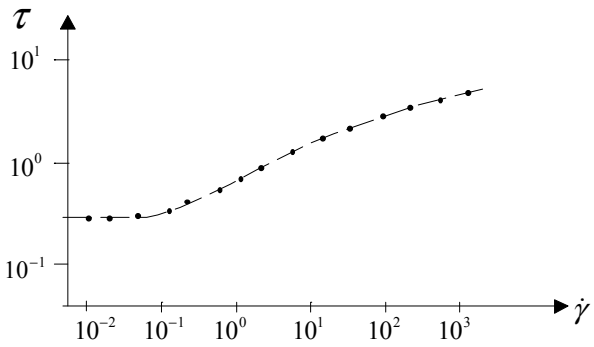


Figure 5.11. *Sketch of a flow curve representative of a yield-stress fluid. Dots represent data and the dashed line the interpolation curve extrapolated towards* $\dot{\gamma} = 0$.

Viscoplastic models

Mostly the *Bingham model* is used to interpolate the data

$$\tau = \tau_c + K\dot{\gamma},$$

with τ_c the yield stress and K a parameter called consistency. A more sophisticated model is the *Herschel-Bulkley model*, which takes nonlinear shearing effects into account

$$\tau = \tau_c + K\dot{\gamma}^n,$$

where $n < 1$ in most cases. Another candidate is the *Casson model*

$$\sqrt{\tau} = \sqrt{\tau_c} + \sqrt{K\dot{\gamma}}.$$

Accurate experimental procedure

Extrapolation of rheometrical data is not recommended to estimate the yield stress. When possible, it is better to use direct tests such as the following trial-and-error procedure:

- a low shear stress is imposed to the sample and the resulting deformation is measured. If there is no yielding, the deformation tends toward a limiting value at long times ;
- when the shear stress imposed is in excess of the yield stress, a constantly rising deformation is observed and the growth rate is the bulk viscosity

The game consists in finding the yield stress by applying successive stress levels to the sample.

Another method proposed by Mendes & Dutra (2004) involves plotting

$$\frac{\mathrm{d}\ln\tau}{\mathrm{d}\ln\dot{\gamma}} = \frac{\dot{\gamma}}{\tau}\frac{\mathrm{d}\tau}{\mathrm{d}\dot{\gamma}},$$

as a function of τ. The yield stress position is given by a sharp peak in the curve. An example is provided in Fig. 5.13.

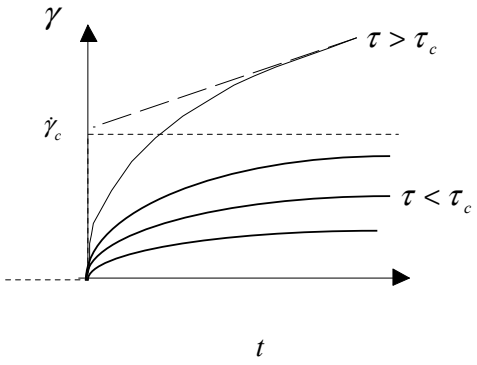


Figure 5.12. *Sketch of the strain variation with time depending on the stress level.*

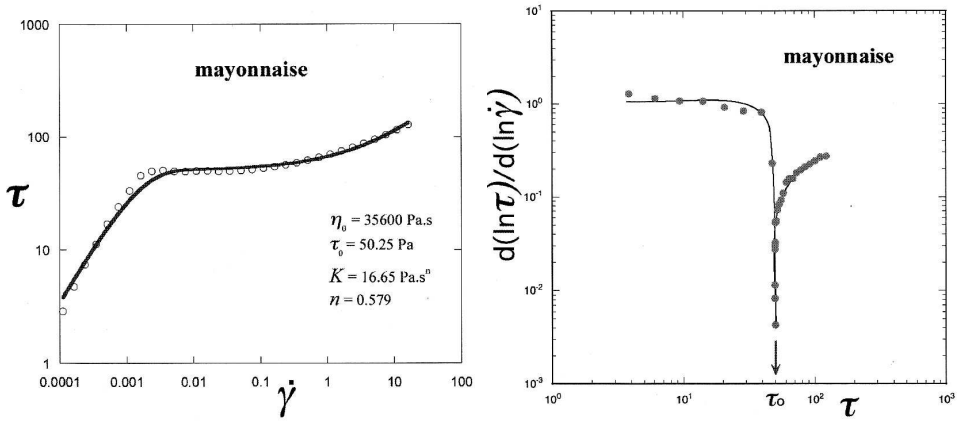


Figure 5.13. *(a) Flow curve of a mayonnaise sample. (b) Estimation of the yield stress. After (Mendes & Dutra, 2004).*

5.5.4 Viscoelasticity

Partitioning materials into fluids and solids is an idealized view. Depending on the typical timescale, a material can be considered as a solid (e.g., exhibiting an elastic behavior) or a fluid. For instance, over 'short' timescales, a glacier is a solid, whereas at long timescales (e.g., several years), it behaves like a liquid. This constat leads many rheologists to state that everything flows, even solids flow, but very slowly. To reconcile this paradoxical absence of differences between what we usually call 'solids' and 'fluids', it is helpful to introduce the notion of relaxation time: the *relaxation time* is the characteristic time needed for a material to flow. In reference to Deborah in the Bible, rheologists also introduce the Deborah number, which is the ratio between the characteristic time/duration of an observation/experiment $t_{obs.}$ and the relaxation time t_r

$$De = \frac{t_{obs.}}{t_r}.$$

When $De \ll 1$, the observer/experimentalist has not the time to register any fluid/creep motion and the material behavior can be considered as solid. When $De \gg 1$, the material has time to relax and modify its structure as a response to the applied forces; the material behaves like a fluid.

Using this definition and using usual timescales for observation/experiments, most ma-

terials belong to either the solid or fluid classes. However, for some materials, the Deborah number is of order of unity, which means that the material can exhibit both solid and fluid properties. Viscoelasticity is a typical trait of materials exhibiting fluid/solid properties.

Linear viscoelasticity

In most textbooks and courses on rheology, the simplest way to introduce the notion of viscoelasticity is to make use of analogies with simple mechanical models consisting of springs (elastic behavior) and dashpot (viscous behavior). These analogues make it possible to have some conceptual insight into the physical behavior of complex materials by breaking down the dissipative viscous processes (time-dependent) and energy-storage processes. Two basic ingredients are used

- spring: according to Hooke's law, the strain γ is proportional to the applied stress τ, which reads
$$\tau = G\gamma,$$
with G the elastic modulus. Note that: (i) for a given stress, there is a limiting deformation $\gamma = G^{-1}\tau$, (ii) the behavior is independent of time. Physically, elastic elements represent the possibility of storing energy. This storage can be achieved by different processes (e.g., polymer recoil).

- dashpot: the response of the dashpot, the plunger of which is pushed at the velocity $\dot{\gamma}$ is
$$\tau = \mu\dot{\gamma},$$
where μ is the viscosity. Note that if we switch on a stress τ, the material response is immediate and the deformation rate is proportional to the applied rate. Physically, dashpots represent dissipative processes that occur as a result of the relative motion between molecules, particles, or polymer chains. This motion induce friction when there is contact between elements or viscous dampening if there is an interstitial fluid.

The simplest representation of a visco-elastic fluid is to combine a spring and a dashpot in series and this combination is called the *Maxwell* model. If the two elements are mounted in parallel, the combination is called a *Kelvin-Voight model* and is the simplest representation of a viscoelastic solid. There are a number of possible combinations of these two elementary models, the simplest one is the *Burgers model*, which is the association of a Maxwell and Kelvin-Voight models. These models are empirical in essence; experimentation showed that these simple models capture the basic properties of a number of viscoelastic materials, but they have their limitations.

Maxwell model. – Let us consider the response given by a Maxwell model: since the spring and the dashpot are in series, the total deformation is the sum of the elementary deformation. We deduce
$$\frac{\mathrm{d}\gamma}{\mathrm{d}t} = \frac{1}{G}\frac{\mathrm{d}\tau}{\mathrm{d}t} + \frac{\tau}{\mu},$$
whose general solution is
$$\tau(t) = Ke^{-\frac{Gt}{\mu}} + \int_{-\infty}^{t} Ge^{\frac{G(t'-t)}{\mu}}\dot{\gamma}(t')\mathrm{d}t',$$

where K is an integration constant, the lower boundary in the integral is arbitrary. If we require that the stress in the fluid is finite at $t = -\infty$, then we must set $K = 0$. Note that:

- for steady state, this equation simplifies to the Newtonian equation $\tau = \mu\dot{\gamma}$;
- for sudden changes in stress, the time derivative dominates;
- the general solution can be cast into the following form

$$\tau(t) = \int_{-\infty}^{t} \left[\frac{\mu}{t_r} e^{\frac{t'-t}{t_r}}\right] \dot{\gamma}(t')\mathrm{d}t' = \int_{-\infty}^{t} \Gamma(t-t')\dot{\gamma}(t')\mathrm{d}t,$$

  where $t_r = \mu/G$ is a relaxation time. The term within the brackets is called the *relaxation modulus* and the integral takes the form of a convolution product of $\Gamma(t) = \mu e^{-t/t_r}/t_r$ and $\dot{\gamma}$. When written in this form, the Maxwell model says that the stress at the present time t depends on the strain rate at t as well as on the strain rate at all past time t' but to within a weighting factor that decays exponentially. This is the simplest representation of *fading memory*. This way of representing the stress is particularly interesting because the integrands is written as the product of two functions: the first one Γ represents the fluid properties, while the second depends on the nature of the flow (via the shear rate). All generalized viscoelastic models are specified in this form.

If we apply this model to the creep testing (see below), the deformation is described by the curve

$$\gamma = \tau\left(\frac{1}{G} + \frac{t}{\mu}\right).$$

Kelvin-Voight model. – The deformation is described by the curve

$$\gamma = \frac{\tau}{G}\left(1 - e^{-\frac{t}{t_r}}\right),$$

where $t_r = \mu/G$ is once again the relaxation time.

Burgers model. – The deformation is described by the curve

$$\gamma = \tau\left(\frac{1}{G_1} + \frac{1}{G_2}\left(1 - e^{-\frac{t}{t_r}} + \frac{t}{\mu_1}\right)\right),$$

where $t_r = \mu_2/G_2$ is the relaxation time.

Creep testing

The simplest test we can imagine is the creeping test: a constant stress is suddenly applied to the material and the strain variation with time is then monitored. The ratio

$$J(t) = \frac{\gamma(t)}{\tau},$$

is usually referred to as the *compliance*.

The typical response can be broken into different phases:

- immediate elastic response;

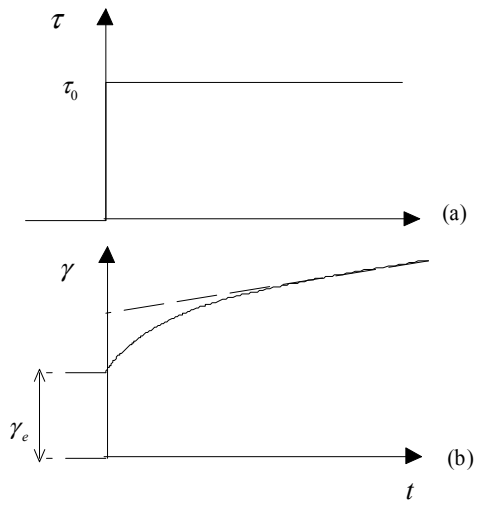


Figure 5.14. *(a) stress variation impose to the sample (b) strain measured as a response to the stress input.*

- delayed elastic response (glassy behavior), where the deformation rate becomes increasingly slow, but ends up as a slow steady-state deformation at long times;
- the steady-state viscous regime, when the material is in steady flow, with constant shear rate (the response curve is a straight line in a plot $\gamma(t)$.

The main features of this common behavior are captured by the Maxwell model (immediate elastic response + steady state behavior) or Burgers model (delayed response is also described).

Oscillatory test

Instead of conducting creeping tests over a range of time, we can run oscillatory shear experiments over a range of frequency. The basic idea is to apply a sine-wave-shaped strain (resp. stress) and measure the resulting stress (resp. strain). For instance, if we impose the deformation (here expressed in a complex form)

$$\gamma(t) = \gamma_0 \Re\left(e^{\imath\omega t}\right),$$

where γ_0 is the strain amplitude (real and positive) and ω the frequency, this means that the shear rate is also imposed

$$\dot{\gamma}(t) = \omega\gamma_0 \Re\left(e^{\imath(\omega t+\pi/2)}\right),$$

where $\dot{\gamma}_0 = \omega\gamma_0$ is the shear rate amplitude.

We measure a material response, which takes the form

$$\tau(t) = A(\omega)\gamma_0 \Re\left(e^{\imath(\omega t+\delta)}\right),$$

where δ is the phase shift, or equivalently

$$\tau(t) = B(\omega)\dot{\gamma}_0 \Re\left(e^{\imath(\omega t-\phi)}\right),$$

relative to the shear rate, with $\phi = \pi/2 - \delta$. For small deformations, the shear stress is assumed to oscillate with the same frequency, but not necessarily in phase:

$$\tau = \Re\left(\tau_0 e^{\imath\omega t}\right),$$

where it is worth noticing that the shear stress amplitude τ_0 is in general complex. It is customary to rewrite the outcome signal by breaking down the in-phase and out-of-phase contributions:

$$\tau_0 = -\imath G^* \gamma_0 \Rightarrow \tau = G'(\omega)\gamma_0 \sin\omega t + G''(\omega)\gamma_0 \cos\omega t,$$

or, in terms of shear rate,

$$\tau_0 = \eta^* \dot{\gamma}_0 \Rightarrow \tau = \eta'(\omega)\dot{\gamma}_0 \cos\omega t + \eta''(\omega)\dot{\gamma}_0 \sin\omega t,$$

with $G^* = G' + \imath G''$ the *complex modulus* and $\eta^* = \eta' - \imath\eta''$ the *complex viscosity.*

It is straightforward to deduce the following relations:

$$G^* = -\imath\omega\eta^*,$$

$$A = \sqrt{G'^2 + G''^2} \text{ and } B = \sqrt{\eta'^2 + \eta''^2},$$

$$\tan\delta = \frac{G''}{G'} \text{ and } \tan\phi = \frac{\eta''}{\eta'}.$$

There is a close correspondence between creeping and oscillatory tests:

- short times correspond to high frequency and usually an elastic response is observed
- behavior at long times is given by low frequencies and is usually of viscous type.

How to interpret the G' and G'' curves? In breaking down the output signal, we introduce functions that are directly related to

- the solid behavior, which is in phase with input signal. The function $G'(\omega) = \omega\eta''$ (in absolute value) is called the *storage modulus* since it reflects relaxation times and elastic modulus. For a perfectly elastic solid, we have $G' = G$ and $G'' = 0$;
- the liquid behavior, which is out of phase (offset of $\pi/2$) relative to the input signal. The function $G''(\omega) = \omega\eta'$ is called the *loss modulus* and characterizes the viscous behavior; it gives information about the (viscous) dissipation in the flow. For a perfectly Newtonian fluid, we have $\eta' = \mu$ and $\eta'' = 0$.

Experimentally we can plot the two curves $G'(\omega)$ and $G''(\omega)$ or equivalently $\eta'(\omega)$ and $\eta''(\omega)$. Sometimes, other quantities characterizing the rheological behavior such as $|G^*|$ or $|\mu^*|$ are also used in graphical representations.

What do we learn from oscillatory tests?

Each community in rheology has its own habits. In polymer science, using oscillatory data is quite common, whereas in the rheology of particle suspension, emphasis is usually given to creeping tests.

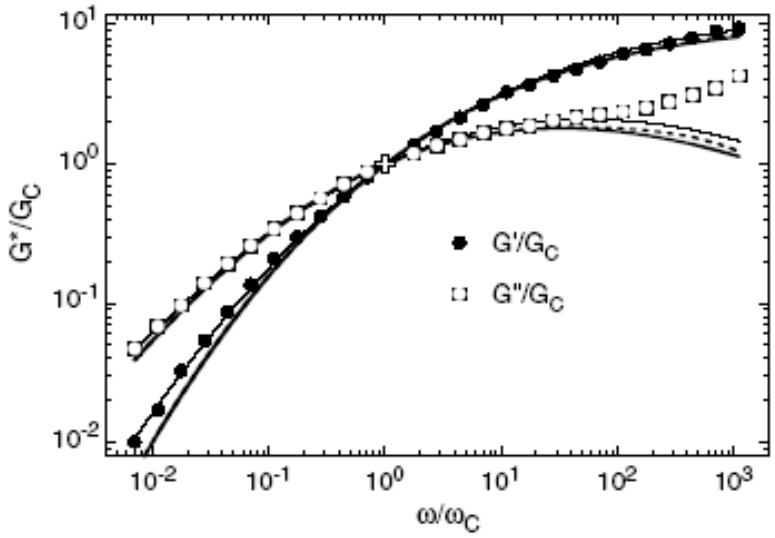


Figure 5.15. *Polysterene dynamic moduli. Points: experimental data from (Guzmán et al.,* Rheol. Acta **44** *(2005) 342 –351). Crossover ($\omega_C = 0.09$ rad/s, $G_C = 23.2$ kPa).*

Figure 5.15 shows a typical example of variations of G' and G''. More information can be inferred from these curves, as shown below.

A number of interesting properties can be pinpointed by looking at G' and G'' curves:

- when $\omega \to 0$ (recall that $\omega \to 0$ corresponds to long time responses in the time domain), $G'' \approx \mu_0 \omega$ since elastic effects are negligible compared to viscous effects at long times, which makes it possible to measure the viscosity at low shear rates. Note that we have always
$$\eta'(\omega)_{\omega \to 0} = \mu(\dot{\gamma})_{\dot{\gamma} \to 0},$$
which implies that $G'' = \eta' \omega$ tends towards 0 and is asymptotic to $\mu(\dot{\gamma})_0 \omega$ (where $\mu(\dot{\gamma})_{\omega \to 0} = \mu(\dot{\gamma})_{\dot{\gamma} \to 0}$).

- We usually observe that the ratio of the storage modulus to ω^2
$$\frac{\eta''}{\omega} = \frac{G'}{\omega^2}$$
tends towards a finite limit when $\omega \to 0$.

- For a number of polymeric liquids (e;g., dilute polymer suspensions), it is quite customary to observe that :
$$\omega \to \infty,\ \eta' \to \eta'_\infty \text{ and } \eta'' \propto \omega^{-1}$$
where η'_∞ is slightly larger than the solvent viscosity. This is not an absolute rule.

- the cross-over frequency at which $G' = G''$ provides an estimate of the longest relaxation time (in s/rad).

- There is an empirical relation referred to as the *Cox-Merz rule*, that relates the viscosity $\mu(\dot{\gamma})$ and the magnitude of the complex viscosity $\eta^*(\omega)$. In short, this rule predicts that

the magnitude $|\eta^*(\omega)|$ is equal to the viscosity at corresponding values of shear rate and frequency

$$|\eta^*(\omega)| = |\eta^*(\omega \to \dot{\gamma})| = \eta'(\omega \to \dot{\gamma})\sqrt{1 + \left(\frac{\eta''}{\eta'}\right)^2}$$

This rule provides a good approximation mostly for polymers. Its use for concentrated suspensions has also been discussed (Doraiswamy *et al.*, 1991; Geissle & Hochstein, 2003).

- There is an equivalent of the Cox-Merz rule for normal stress, which is called *Laun's rule*. This rules states that the material function $\Psi_1 = N_1/\dot{\gamma}^2$ is well approximated by

$$\Psi_1 = \frac{N_1}{\dot{\gamma}^2} = \frac{2\eta''(\omega)}{\omega}\left[1 + \left(\frac{\eta''}{\eta'}\right)^2\right]^{7/10},$$

by replacing ω with $\dot{\gamma}$. This rule originates in the behavior of G'' when $\omega \to 0$ (see above).

Figure 5.16. *Weissenberg effect. Photo from Gareth McKinley's group at MIT.*

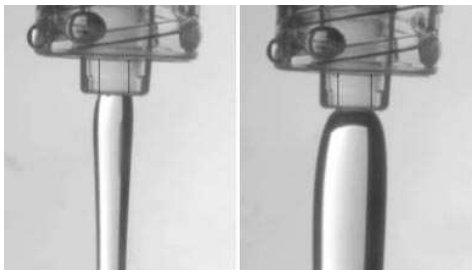


Figure 5.17. *Die swell effect. Photo from Gareth McKinley's group at MIT.*

5.5.5 Normal stress effects

Manifestation

There are many situations where normal-stress effects occur and give rise to specific phenomena. Note that these effects are typical for non-Newtonian (e.g., viscoelastic fluid) fluids. Here are some typical examples:

- Weissenberg effect or rod-climbing: when rotating in a fluid contained in a container, a cylinder disturbs the free surface shape differently depending on whether the fluid is Newtonian or not. For a Newtonian, centrifugal forces induces dipping (fluid expelled from the cylinder vicinity), whereas for elastic fluids, the fluid climbs along the cylinder (see figure 5.16).
- Die-swell effect: when an elastic fluid flows out of a tube, it swells as a result of normal stress effects, whereas for a Newtonian fluid, there is usually a contraction or the surface remains straight.

A number of odd phenomena are also observed in the development of instabilities in elastic fluids (Shaqfeh, 1996; Boger & Walters, 1993).

Origin

Normal stress effects are common with polymeric fluids and in this case, they are usually caused by the relaxation of polymer coils, which had been extended earlier. For instance, in a simple shear flow such as a flow through a die, the polymers are stretched along an elongational axis (direction of the flow), which creates additional tension in this direction

due to chain elasticity. This extra tension is proportional to the elongation, which in turn is proportional to the shear rate $\dot{\gamma}$ and since the normal stress for elastic polymers is dependent on both the elongation and shear rate, we expect a quadratic dependence of the normal stress difference on the shear rate. This scaling is observed experimentally (Bird *et al.*, 1987).

Other materials such as particle suspensions may exhibit non-zero stress differences. In that case, the normal stress effect reflects a nonsymmetric stress field due to an anisotropic structure and/or complex interactions. For instance, a compacted granular medium exhibits clearly normal stress differences due to nonsymmetric force distribution within a particle network (e.g., see Fig. **??** in Chap. **??**) (Wieghardt, 1975). Another example is provided by suspensions of noncolloidal particles, which exhibit specific particle arrangements that result in the normal stress differences (Brady & Morris, 1997; Zarraga *et al.*, 2000).

Measurement

The normal stresses or normal stress differences can be measured in a number of ways. For instance, we have seen that when a fluid flows down an inclined channel, its free surface can be slightly convex because of the normal stress difference (if the tension surface is negligible) and measuring the shape of the free surface in the cross-stream direction makes it possible to evaluate the first normal stress difference (see also pp. 102–105, Tanner, 1988). Similarly, the flow-depth profile in the rod-climbing experiment or the die swelling when extruding polymers can be used to measure normal stresses. In the latter case, Tanner (1988) shows that the ratio of the jet diameter (D_j) and the capillary diameter (D) is connected to the following stress combinations via the approximate relation

$$\frac{D_j}{D} = 0.1 + \left(1 + \frac{1}{2}\left(\frac{\sigma_{zz} - \sigma_{rr}}{2\tau}\right)^2\right),$$

for elastic polymers.

In practice, the cone-and-plate geometry is commonly used to measure the first normal stress difference. We can express the first normal coefficient as

$$\Psi_1 = \frac{2}{\pi}\frac{F}{R^2\omega^2}\varphi^2,$$

where the thrust exerted by the fluid on the cone is denoted by F, the cone angle is denoted by φ, its diameter is R, and ω is the rotational velocity of the geometry.

5.5.6 Thixotropy

Definition

Thixotropy[5] refers to the property of a material such that

- when it flows, its rheological properties vary strongly with shear rate (considerable shear thinning);
- when it is left at stand (after flow cessation), it retrieves its initial consistency (solid-like behavior) and properties.

[5] Coined from the Greek words $\tau\eta\iota\chi\iota\sigma$ (stirring or shaking) and $\tau\rho\epsilon\pi o$ (turning or changing).

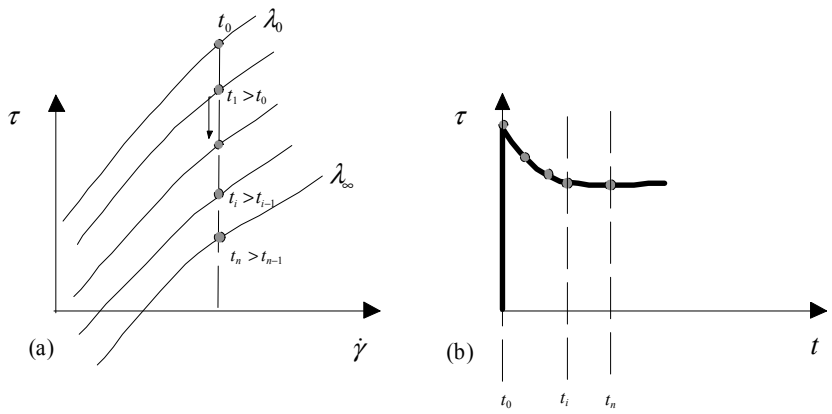


Figure 5.18. *Time decrease in shear stress: because the flow curve actually depends on the microstructure (represented here via a microstructural parameter (a), then the shear stress varies with time before eventually its steady-state value (b).*

Note that there is no strict consensus within the rheologist community concerning this definition (Barnes, 1997). For a number of people (especially in industry), thixotropy means extreme shear thinning and conferring this property to manufactured slurries such as paints is highly desirable.

Manifestation

In 1923, Schalek and Szegvari found that aqueous iron-oxide "gels have the remarkable property of becoming completely liquid through gentle shaking alone, to such an extent that the liquified gel is hardly distinguishable from the original sol. These sols were liquified by shaking, solidified again after a period of time... the change of state process could be repeated a number of times without any visible change in the system" (cited in Barnes, 1997). It was believed that a new kind of phase change had been found. The first manifestation of thixotropic materials is thus the fluidization of the material, a kind of yoghurt effect: when one vigorously mix a yoghurt with a spoon, the consistency alters quite abruptly.

There are several clues that can lead to thinking that a material is thixotropic:

- If we apply a constant shear rate to thixotropic material, the shear stress measured (or, equivalently, viscosity) decreases with time and it will eventually reach a limiting steady value. First this is an overshoot in the shear-stress response, then a slow decrease [see Fig. 5.18(b)]. However, the peak value depends on how carefully or vigorously the material was initially loaded into the rheometer and how long it was left to rest before shearing.

- When applying an increasing shear rate to a material, then at a given by reversing the shear rate [see Fig. 5.19(a)], one can observe a loop in the time record of stress. Some rheometers propose a special function for measuring thixotropy (by measuring the loop area), but in fact, the actual loop area depends on a number of parameters (duration of each shear-rate increment, buildup rate, etc.), which makes it difficult to propose a proper interpretation of this test.

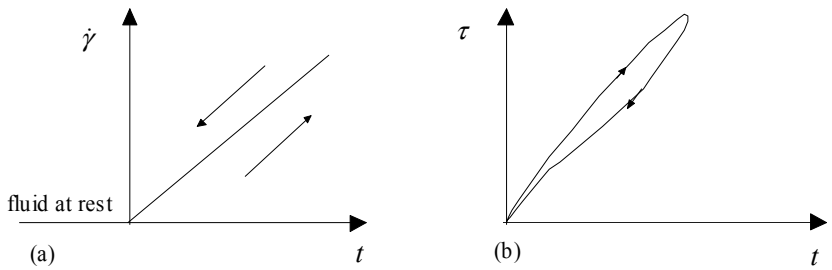


Figure 5.19. *Thixotropy loop.*

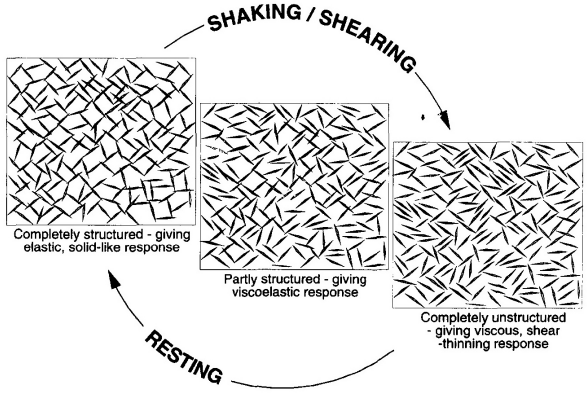


Figure 5.20. *Thixotropy as the manifestation of microstructure influence on bulk rheological behavior. After Barnes (1997).*

Physical origin

Almost all materials that are shear thinning are thixotropic because it takes a finite time to bring about the rearrangements needed in the microstructural elements that result in shear thinning. Typical examples include clays and soil suspensions, creams, drilling muds, flour suspensions, fibre greases, jellies, paints, etc. Three ingredients are usually required to observe a thixotropic behavior (see 5.20):

- a material made up of "structures" that progressively break down on shearing (or shaking);
- a reversible process that implies that the "structuration" of the material lost after flow inception is retrieved when the material has been left at rest for a sufficiently long time. Thus, shearing induces loss of the initial order (at rest), while resting implies rebuilding;
- the timescales characterizing each phase (structure breakdown/rebuiling) are not equal, with usually a characteristic time ranging from a few milli-seconds to a few minutes for the structure breakdown, whereas the rebuilding timescale is of order of a few hours to several days.

Indeed, when at rest, a material possesses a structure that maximalizes the shear strength (both for viscous and elastic processes): there is no order in the spatial distribution of particles and the number of contacts between particles is large. On shearing, particles start aligning, the number of contact between neighboring particles decreases, the particle spatial distribution is asymmetrical in the flow direction.

Characterization

▲ Be careful and remind that some disturbing effects (e.g., slipping) can produce effects similar to thixotropic effects.

▲ A thixotropic material is highly dependent on its past history, especially the deformation history. In theory, if we wait a sufficiently long time, a material at rest has recovered its initial state (e.g., random structure). In practice, one can define an arbitrary initial state (which does not necessarily correspond to a resting/equilibrium state) by applying high shear rates to complete break down the inner structure, then leaving the material at stand for a given time. Proceeding with the material in this way makes it possible to have the same initial state as many often as desired.

▲ When testing the rheological response of a material, it can be recommended to test it within a homogenous flow, i.e. within a flow where the shear rate is constant and uniform at any point and time. In the converse case, the state of the material being dependent on shear-rate, its local response may differ significantly depending on the position in the flow. Let us take the example of a parallel-plate rheometer, we have $\dot{\gamma} = 0$ at $r = 0$, which implies weak breakdown in the vicinity of the central axis, whereas at the periphery, the material is fully disorganized.

We suggest the following procedure:

1. Let us assume that we start from a well-specified initial state, the material being at rest. We apply a constant shear rate $\dot{\gamma}_0$ [see Fig. 5.21(a)] and measure the shear stress.

2. On applying a shear rate, the inner structure is disorganized, which requires energy. This is reflected by a shear-stress overshoot, which is followed by a slow decrease towards a limiting value corresponding to the new equilibrium induced by the new strain rate [see Fig. 5.21(b)] if we wait a sufficiently long time.

3. If we remove the shear rate, the shear stress drops to zero instantaneously (no viscoelasticity).

4. After a resting time, we apply the same rate as earlier, but since the inner structure was altered, less energy is needed to break the particle network or realign the particles, which implies that the shear-stress overshoot is less pronounced and the steady state is reached more quickly [see Fig. 5.21(b)].

5. If the overshoot is denoted by $\Delta\tau$, we can determine the function $\Delta\tau(t_{rest}$ relating the overshoot to the resting

Modelling

The simplest mode we can imagine is to assume that the viscosity is dependent on a structural parameter. As a first approximation and for the sake of simplicity, we assume that (i) this parameter is scalar, (ii) it reflects the microstructure state (e.g., the floc size in a flocculated suspension, the number of contact in a colloid, available energy for a given particle configuration, etc.), (iii) it satisfies a kinetic equation in the form

$$\frac{\mathrm{d}\lambda}{\mathrm{d}t} = \frac{F(\lambda)}{t_b} - \frac{G(\lambda)}{t_e},$$

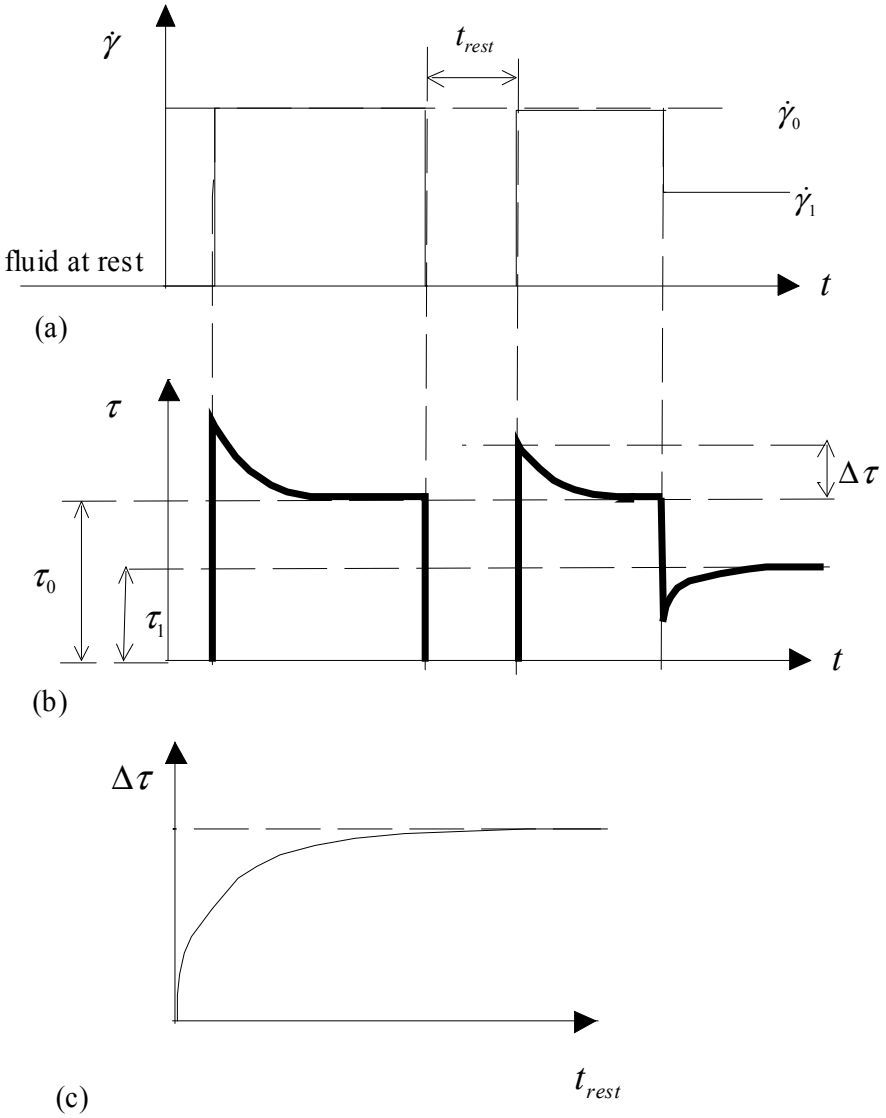


Figure 5.21. *Typical evolution of shear stress (b) in response to a shear-rate history (a). Iterating the procedure with different values of shear-rate levels leads to determining the stress jump $\Delta\tau$ as a function of the resting time t_{rest}.*

where t_b is the buildup timescale, $t_e = \dot{\gamma}^{-1}$ is a typical characteristic of the flow, F represents the increase of the scalar parameter induced by buildup, whereas G stands for the λ as a result of microstructure disorganization. These functions can de determined experimentally. Typical examples include $F = (1 - \lambda)^a$ and $G = b\lambda$, with a and b two reals (see pp. 24–25, Barnes, 1997)

In short, we have $\mu = \mu(\dot{\gamma}, \lambda)$, with λ variations controlled by the kinetic equation above. When solving a problem, we have to solve the momentum equations together with the kinetic equation.

5.6 Problems encountered in rheometry

5.6.1 Problems with rheometers

In rheometry, many disturbing effects may arise. They often reflect the influence of the microstructure. For instance, for a particle suspension, especially made up of nonbuoyant particles, sedimentation and migration of particles can significantly alter the stress distribution and thus the measured torque. Likewise, for concentrated pastes, a fracture inside the sheared sample may sometimes be observed, usually resulting from a *localization of shear* within a thin layer. Other disturbing effects are experimental problems pertaining to the rheometer type. For instance, when using a rotational viscometer with a smooth metallic shearing surface, wall slip can occur. Apart from effects resulting from microstructural changes, which are a part of the problem to study, it is sometimes possible to reduce disturbing effects or to account for them in the flow-curve computation.

Problems coming from electronic defects

As for any electronic apparatus, a rheometer may encounter problems that cause trouble when taking measurement. The typical sources of error include

- calibration defect: test the response of the rheometer by using a previously calibrated fluid, usually stable oil (e.g., silicone). These may be purchased from chemical product suppliers, who propose ISO-9002 registered and NIST-traceable fluids.
- zero drift: test the rheometer with a Newtonian fluid.

End effects

More important are probably the problems related to wall and end effects. Any rheometer is subjected to end effects, which have to be corrected or taken into account in the computation of the flow curve. Typical examples are the following

- end effects in a channel are due to the finite length of the channel as well as the sidewalls, both producing potentially significant variations in the flow depth;
- in a Couette rheometer, the measured torque includes a contribution due to the shearing over the bottom surface of the bob. Such a contribution is substantially reduced using a bob with a hole hollowed on the bottom surface so that air is trapped when the bob is immersed in the fluid. But this can be inefficient for particle suspensions, such as granular materials, and in this case, the bottom contribution to the resulting torque must be directly assessed using the method proposed by Barnes & Carnali (1990);
- for a parallel-plate rheometer, the fluid surface at the peripheral free surface may bulge out or creep, inducing a significant variation in the measured torque, possibly varying with time.

Wall effects: slipping and adherence

A substantial source of problems arises with particle suspensions due to the presence of a wall (Barnes, 1995):

- a wall modifies the particle arrangement: this phenomenon called *particle depletion* involves a decrease in particle concentration close to the wall, which leads to the development of a lubricated fluid layer close to the solid boundary and to the slipping of the bulk.
- depending on the fluid properties, there may be interactions between the metallic surface and the constituents of the liquid. For specific materials like flocculated suspensions (floc size is shear-rate dependent) or for fibrous materials, the presence of a solid boundary may alter the local structure within the liquid, thus local viscosity.

Slipping problems can be pinpointed experimentally by marking the sample with very fine non-active powder at the free surface and on the edges of the cone and plate, as shown in figure 5.22(a). Slipping and sample rupture is detected by observing how lines deform during the test. When the deformation is homogeneous, as expected or desired, the marked lines are straight, as shown in 5.22(b). Slipping is made visible through discontinuities of the marks.

Slipping may significantly disturb measurements :

- under-evaluation of bulk viscosity
- improper evaluation of the yield stress for a viscoplastic material

Some solutions are:

- A strategy involves measuring the slipping velocity directly and then computing an effective shear rate.
- Still another possibility requires using the same rheometer with different sizes, as first proposed by Mooney for the capillary rheometer.
- Rheometer suppliers provide specific grooved or corrugated geometries. Sandblasting with a coarse grit or gluing a sand paper can also be used to roughen a metallic surface.
- A growing number of applications are based on vane shear cell Barnes & Nguyen (2001).
- When there are chemical interactions (chemical attack with ion production) or physical interactions (van der Waals force) between the fluid constituents and the walls, specific surface treatment must be used.

Be careful: some techniques such as using rough surfaces do not remove the slipping problem, but only shift it: instead of slipping, the material undergoes shear localization within the sample.

Rupture within the sample

With rupture, we mean:

- shear localization: shear is concentrated within a narrow strip inside the sample, which implies a jump in shear rate on the bulk scale

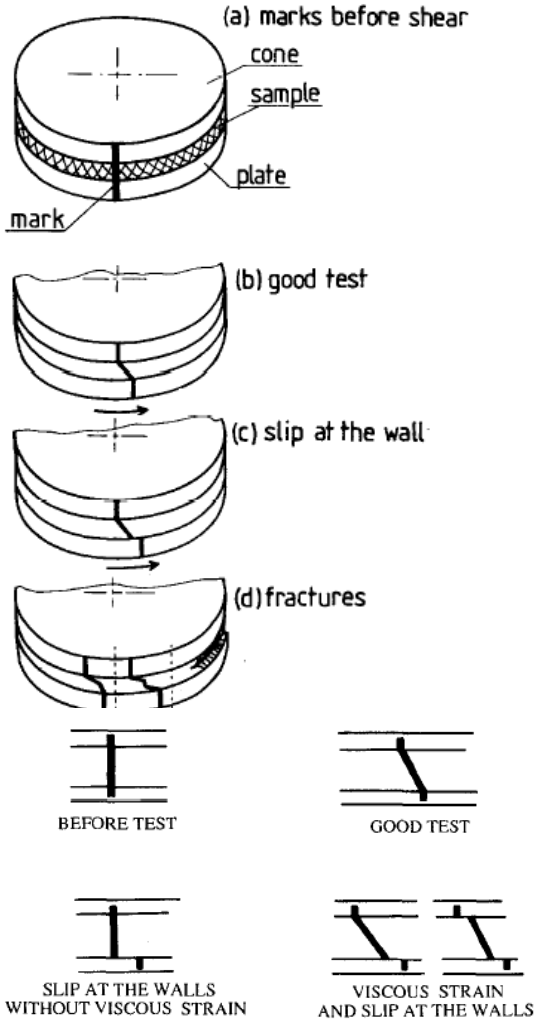


Figure 5.22. *Detection of slipping and shear localization for shear tests. After (Magnin & Piau, 1987, 1990).*

- slipping between two layers of material.

Deformation of the free surface

In most viscometric flows, there is a free surface. This surface may deform as a result of different processes:

- normal stress effect: rod-climbing effect or bulging of the free surface for flows down inclines
- inertial effect: for the coaxial cylinder rheometer at high velocities, there is a gradient in the flow depth to balance the inertial centrifuge forces. For the parallel-plate rheometer, fluid may be expelled because of radial acceleration
- evaporation, particle migration etc. at the periphery of the sample for parallel-plate rheometer.

Finite-size effects

Furthermore, many natural fluids encountered in geophysics are suspensions with a large size distribution. The size of the rheometer should be determined such that its typical size (e.g., the gap in a rotational viscometer) is much larger than the largest particle size.

The solution involves using large-sized rheometers, e.g., wide gap Couette cells.

5.6.2 Limitations of the viscometric treatment

The 'simple-fluid' assumption

The viscometric treatment relies on the crucial assumption that the extra-stress tensor is a one-to-one function of the strain-rate tensor only (class of simple fluids). Many classes of material studied in geophysics are not in fact incompressible, homogeneous, isotropic, or merely expressible in the form $\boldsymbol{\sigma} = -p\mathbf{1} + \mathbf{s}(\mathbf{d})$.

For instance, for materials with time-dependent properties (thixotropic materials, viscoelastic materials), the constitutive equation can be expressed in the form $\boldsymbol{\sigma} = -p\mathbf{1} + \mathbf{s}(\mathbf{d})$ only for a steady state. Another example is provided by granular flows. In this case, when applied to experimental data obtained by studying dry granular flows down an inclined channel (Ancey *et al.*, 1996), the viscometric treatment leads to the conclusion that the flow curve should be a decreasing function of the shear rate in violation of a stability criterion imposing that the flow curve be an increasing function. Although such a decrease in the flow curve cannot be directly interpreted in terms of a constitutive equation, it provides interesting rheological information that can be explained on the basis of microstructural theories (Ancey & Evesque, 2000).

Development of instabilities

At high speed, a number of instabilities occur:

- development of secondary flow and recirculation cell, e.g., Saffman-Taylor instability in

Couette cells

- turbulence development
- particle migration, particle jamming or settling

The possible solution is to

- visualize the internal flow structure to detect secondary flow;
- place tracers in the fluid to determine their trajectory, then the streamlines;
- check the consistency of the data. For instance for a parallel-plate rheometer, the viscometric treatment is valid provided centrifugal forces are negligible compared to the second normal stress difference: $\varrho R^2 w^2 \ll N_2$, where w is the orthoradial component of the velocity. Such an effect can be detected experimentally either by observing secondary flows or by noticing that doubling both the gap and the rotational velocity (thus keeping the shear rate constant) produces a significant variation in the measured torque.

5.6.3 Technical issues related to the derivation of the flow curve

Recall that:

- The softwares provided by rheometer manufacturers for computing the flow curve or other rheological properties makes use of standard methods. For instance, for the Couette cell (coaxial cylinder rheometer), the narrow-gap approximation is used, which implies that the user should develop his own software to deal with wide-gap rheometer.
- For most viscometers, computing the shear rate from experimental data can raise serious problems. A major source of uncertainty is that in most viscometric procedures the shear rate is expressed as a derivative – for instance $\partial q/\partial h$ in (5.11) – which must be estimated from experimental data. To do so, different procedures are available but they do not always provide the same results, especially when data are noisy (Borgia & Spera, 1990).

A typical example of these problems is given by the concentric-cylinder rheometer (or Couette rheometer). The shear rate is inferred from the rotational velocity Ω and the torque (per unit depth) C using the following relationship:

$$\Omega = -\frac{1}{2}\int_{C/(2\pi R_1^2)}^{C/(2\pi R_2^2)} \dot{\gamma}(\tau)\mathrm{d}\ln\tau. \tag{5.18}$$

When the gap between the two cylinders is narrow, it is possible to approximate the shear rate as: $\dot{\gamma} = R_1\Omega/(R_2 - R_1) + o(1 - R_2/R_1)$.

However, such a geometry is not very suitable to studying particle suspensions (slipping, size effects, etc.) and usually a wide gap is preferred. For methods of this kind, computing the shear rate requires:

- specifying the type of constitutive equation in advance, integrating it to obtain the relation between the torque and the rotational velocity, and fitting the relation to experimental data.

- using a more effective and practical method of solving the inverse problem such as
  - the Tikhonov regularization method proposed by Yeow *et al.* (2000): this procedure does not require the algebraic form of the $\tau - \dot{\gamma}$ curve to be specified and has the advantage of filtering out noise.
  - the wavelet-vaguelette decomposition method proposed by Ancey (2005) which is not based on a regularization step and thus is more appropriate for complex fluids.

The only viscometer which poses no problem in converting experimental data into a $\tau - \dot{\gamma}$ curve is the parallel-plate rheometer. In this case, the shear rate distribution is imposed by the experimentalist: $\dot{\gamma} = \Omega R/h$. But such a relationship holds provided centrifugal forces are negligible compared to the second normal stress difference: $\varrho R^2 w^2 \ll N_2$, where w is the orthoradial component of the velocity. Such an effect can be detected experimentally either by observing secondary flows or by noticing that doubling both the gap and the rotational velocity (thus keeping the shear rate constant) produces a significant variation in the measured torque.

5.6.4 Problems related to sample preparation

A sample that will be tested in a rheometer must be representative of large batches of material from which it is extracted. Specific care should be taken in obtaining and preparing particle suspensions when the sample has been collected in the field (e.g., a sample of debris flow) or in industrial facilities due to segregation effects.

5.7 Non-standard techniques: what can be done without a rheometer?

5.7.1 Viscosity: free fall of a bead

Consider an isolated spherical particule of radius a falling under the action of gravity in a fluid that is assumed to be Newtonian with viscosity μ and unbounded (no wall close to the particle). If the sphere moves very slowly, its Reynolds number

$$\mathrm{Re} = \frac{\varrho_p a u}{\mu},$$

with ϱ_p the particle density and u the particle velocity relative to the fluid. When $\mathrm{Re} \ll 1$, then the force exerted by the fluid on the particle is given by Stokes' law

$$F = 6\pi\mu a u.$$

If we can measure the velocity u when the particle reaches a steady regime, we can determine the fluid viscosity from the force balance $0 = -\varrho_p g + \varrho_f g + F$, where fluid density has been denoted by ϱ_f

$$\mu = \frac{2}{9}(\varrho_p - \varrho_f)\frac{g a^2}{u}.$$

A practical version of this measurement technique is the rolling-ball viscometer (Bird *et al.*, 1987).

5.7.2 Yield stress: Slump test

In the laboratory, it is frequently impossible to investigate the rheological properties of a particle suspension using a rheometer. For instance, with snow or magma, such tests are almost always impractical. For debris suspensions, it is usually impossible to carry out measurements with the complete range of particle size. This has motivated researchers to developed approximate rheometric procedures and to investigate the relations between field observations and rheological properties. For instance, given the sole objective of determining the yield stress, the semi-empirical method referred to as a slump test can provide an estimate of the yield stress for a viscoplastic material. This method involves filling a cylinder with the material to be tested, lifting the cylinder off and allowing the material to flow under its own weight. The profile of the final mound of material as well as the difference (δ) between the initial and final heights is linked to the yield stress. Pashias & Boger (1996) have found:

$$\frac{\delta}{h} = 1 - 2\frac{\tau_c}{\varrho g h}\left(1 - \ln\left(2\frac{\tau_c}{\varrho g h}\right)\right), \tag{5.19}$$

where h is the cylinder height, ϱ the material density (see also Schowalter & Christensen, 1999). Close examination of experimental data published by Pashias and Boger shows a deviation from the theoretical curve for yield stress values in excess of approximately $0.15\varrho g h$. For yield stress values lower than $0.15\varrho g h$ (or for $\delta/h > 0.4$), uncertainty was less than 10% for their tests. The explanation of the deviation for higher yield stress values lies perhaps in the weakness of the assumption on the elastoplastic behavior for very cohesive materials. **?** developed an alternative approach based on an interpretation of the deposit shape. They

showed that the free surface profile (the relationship between the material height y and the distance from the edge x) depends on the yield stress only. On a flat horizontal surface, the free surface profile has the following expression

$$\frac{\varrho g y}{\tau_c} = \sqrt{2\frac{\varrho g x}{\tau_c}}. \tag{5.20}$$

Comparisons between rheological data deduced from a parallel-plate rheometer and free surface profile measurements showed an acceptable agreement for fine mud suspensions and debris flow materials. Uncertainty was less than 20%, within the boundaries of acceptable uncertainty for rheometrical measurement. The major restriction in the use of equation (5.20) stems from the long-wave approximation, which implies that the mound height must far outweigh the extension of the deposit: $h - \delta \gg \tau_c/(\varrho g)$. The method proposed by **?** can be extended to different rheologies and boundary conditions. In the field, such a method applied to levee profiles of debris flow can provide estimates of the bulk yield stress provided that the assumption of viscoplastic behavior holds.

A number of other tests have been developed in civil engineering tests (e.g., Abram's test for evaluating concrete workability) and in the industry (e.g., Bostwick's test in food engineering), which usually provide information on how easily a material flows. More recently, some theoretical analyses have tried to elaborate on the interpretations that can be done with these empirical (McKarthy & Seymour, 1994; Piau & Debiane, 2005).

A Phase portrait

A.1 Introduction

A number of nonlinear equations of first order as well as second-order autonomous equations can be cast in the following form:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{f(x,\ y)}{g(x,\ y)}, \tag{A.1}$$

with f and g two functions that may vanish. The points that are both zeros of f and g are called *singular points* [1] since the differential term $\mathrm{d}y/\mathrm{d}x$ is *a priori* indeterminate at these points. The behavior of the integral curves depends strongly on the structure of curves $f(x,\ y) = 0$ and $g(x,\ y) = 0$ around these critical points, i.e. the multiplicity of critical curves generated by the equations $f(x,\ y) = 0$ and $g(x,\ y) = 0$ and by the sign of f/g in the different areas delineated by these critical curves.

The simplest case is encountered when, near the singularity, it is possible to linearize Equation (A.1). We can then write: $f(x,\ y) = ax+by+o(x,\ y)$ et $g(x,\ y) = cx+dy+o(x,\ y)$. Let us assume that $ad - bc \neq 0$ and these coefficients are not all zero. There are two critical curves in the vicinity of the singularity:

- $y = -ax/b$ where the curves admit a horizontal tangent;
- $y = -cx/d$ where the curves admit a vertical tangent.

Introducing dummy variable t, we can transform (A.1) into the matrix form:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u} = \mathbf{M}\cdot\mathbf{u}, \text{ with } \mathbf{M} = \begin{bmatrix} c & d \\ a & b \end{bmatrix}. \tag{A.2}$$

We seek a solution in the form $\mathbf{v} = \mathbf{v}_0 \exp(\lambda t)$, with $\mathbf{v}_0$ the initial-condition vector (at $t = 0$). We deduce that λ must be an eigenvalue of the matrix $\mathbf{M}$ and $\mathbf{v}_0$ an associate eigenvector; λ is solution to the second order equation $\lambda^2 - 2h\lambda + k = 0$, with $2h = b + c$ and $k = \det \mathbf{M} = -ad + bc$, that is:

$$\lambda = h \pm \sqrt{h^2 - k}.$$

[1] They are also called *critical* points or *equilibrium* points.

The principal directions principales are: $b - c \mp \sqrt{(b-c)^2 + 4ad}$. Depending on the value taken by λ, different behaviors arise:

- when $\Delta = h^2 - k > 0$ and $k > 0$, the two eigenvalues are real and of the same sign. Assume that $h > 0$, then the two eigenvalues are positive, which means that either both solutions to (A.2) tend to 0 as $t \to -\infty$ (resp. when $h < 0$, the solutions tend to 0 as $t \to +\infty$). Hence, if every initial condition lies on one of the principal axes, each solution tends to the origin point, the solution is a part of a straight line with slope equal to one of the principal directions. What happens if the initial condition does not lie on one of the principal directions? Let us assume that an integral curve tends towards the origin point. The limit of $\mathrm{d}y/\mathrm{d}x$ at 0 in the equation (A.1) is not defined. Applying Rule's Hospital (see below), the slope of the solution at point O must satisfy:

  $$m = \frac{a + bm}{c + dm},$$

  i.e., $m = b - c \pm \sqrt{(b-c)^2 + 4ad}$ and m coincide with one of the main directions. Given the sign of $\mathrm{d}y/\mathrm{d}x$ around the origin point, only one of these solutions is possible: the curves reach the origin point, following an asymptotic curve of equation $y = mx$. This singularity is called a *node*. Figure A.1 shows an example.

- if $\Delta > 0$ and $k > 0$, the two eigenvalues are real and of opposite sign. The two solutions of system (A.2) behave differently when $t \to \infty$: one tends towards the singular point while the other tends to infinity. There are always two curves that pass through the singular point and that coincide with the principal directions. If now the initial point (initial condition of the differential equation) does not lie on one of the principal directions, then it is not possible to find an integral curve emanating from that point to the singular point because of the sign of $\mathrm{d}y/\mathrm{d}x$ in the close vicinity of the singular point. The paths diverge when approaching the singular point. We refer to this point as a *saddle*. Figure A.2 shows an example.

- if $\Delta = 0$, the singular point is a node.

- if $\Delta < 0$, both eigenvalues are imaginary. The curves coiled like a spiral around the singular point. This point is called *focal point*. Figure A.3 shows an example.

♣ **Example**. – When solving the equation:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{x + 2y}{2x + y}, \tag{A.3}$$

we find out that there are two eigenvalues 3 and 1 associated with principal directions 1 and et -1 respectively. It is a node.

♣ **Example**. – When solving the equation:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{2x + y}{x + 2y}, \tag{A.4}$$

we find out that there are two eigenvalues 3 and -1 associated with principal directions 1 and et -1 respectively. It is a saddle.

♣ **Example**. – When solving the equation:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{2x+y}{x-y}, \tag{A.5}$$

we find out that there are two complex eigenvalues $(3 \pm \imath\sqrt{3})/2$. It is a focal point.

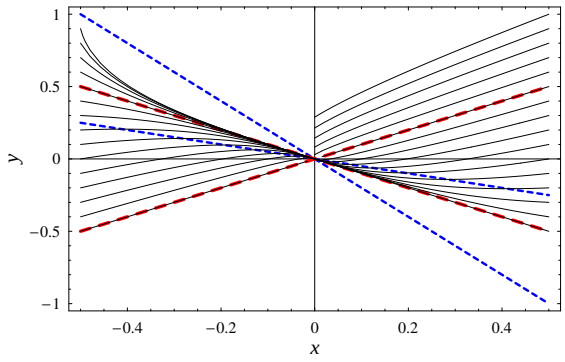


Figure A.1. *Example of node. Solid lines are solutions to the differential equation (A.3). Blue dashed lines represent singular curves while the red lines stand for the principal directions.*

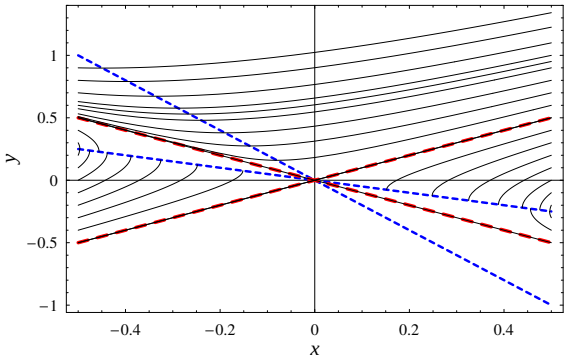


Figure A.2. *Example of saddle. Solid lines are solutions to the differential equation (A.4). Blue dashed lines represent singular curves while the red lines stand for the principal directions.*

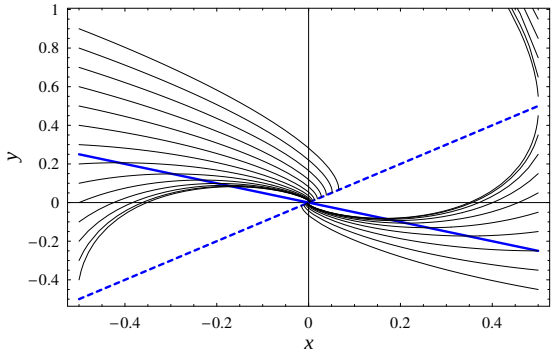


Figure A.3. *Example of focal point. Solid lines are solutions to the differential equation (A.5). Blue dashed lines represent singular curves.*

A.2 Typology of singular points

The preceding discussion can be generalized to forms of differential equations that are much more complex than the linear system (A.2). We note that there are three possible types of behavior:

- *node* where the integral curves (there exists an infinity of such curves) are directed towards the singular point, usually following an asymptotic curve that can be deduced from the differential equation;

- *saddle* where the integral curves diverge as they approach the singular point, except one who is able to cross it;
- *focal point* where the curves wrap up like spirals or make loop around the singular point.

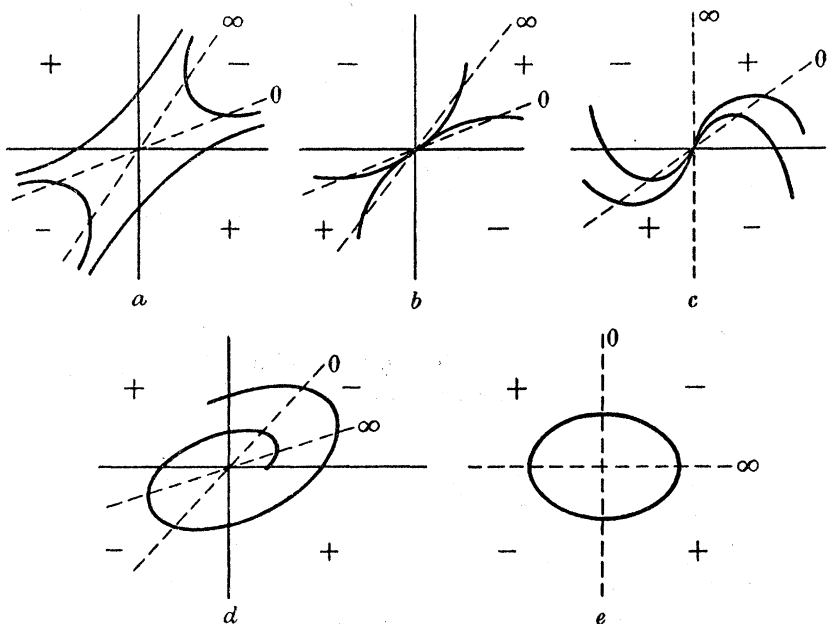


Figure A.4. *Typology of singular points where there are two critical curves. After (Jones, 1953).*

In the general case, the behavior of functions is a combination of these three basic forms, which is more or less complex depending on the number of critical curves $f = 0$ and $g = 0$. Figure A.4 recalls the possible behaviors when there are two critical curves. Figure A.5 shows the possible combinations when there are three critical curves (two corresponding to $f = 0$ et a single one to $g = 0$). Figure A.6 shows the possible combinations when four critical curves, two of which coincide with the coordinate axes.

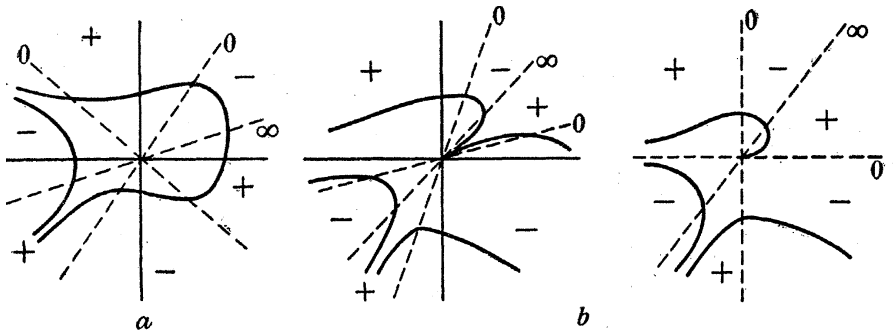


Figure A.5. *Typology of singular points where there are three critical curves. After (Jones, 1953).*

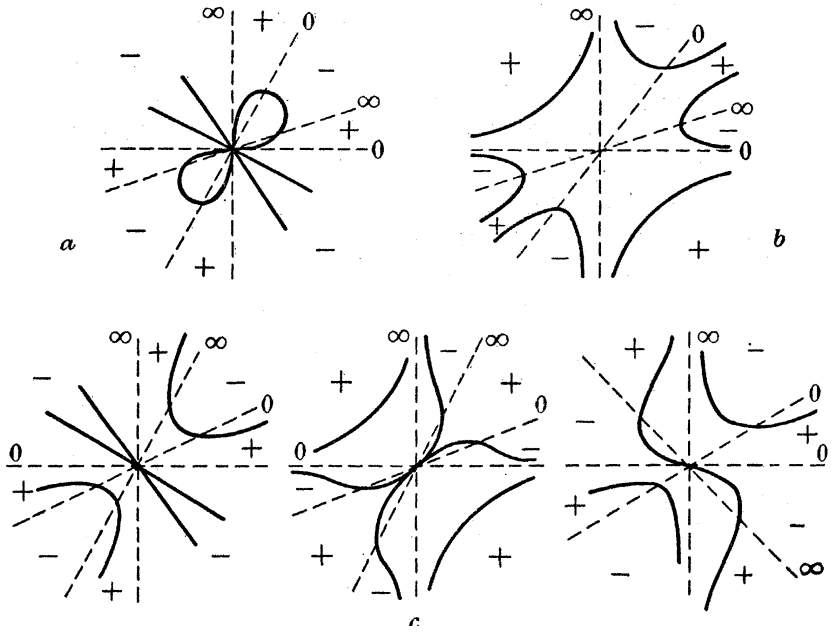


Figure A.6. *Typology of singular points where there are four critical curves. After (Jones, 1953).*

A.3 Computation of the asymptotic curve

When the singularity is a node, there is an asymptotic curve to which any curve passing through the singularity tends. Similarly, when the singular point is a saddle, there is a (single) curve solution that arrives at the singular point. The exceptional curve is called *separatrix*. because it also separates two regions of space, each characterized by a specific behavior near the singular point. We can use several methods to work out the equation of this curve.

A.3.1 Numerical computation

Using l'Hôpital's rule, we can obtain the asymptotic curve toward which the integral curves moving towards a node converge or the single curve passing through a saddle. Indeed we can write:

$$F(x) = f(x,\ y(x)) \quad \text{and} \quad G(x) = g(x,\ y(x)).$$

By making a first-order expansion around a singular point $\mathbf{x}_s$, we have:

$$\dot{y}_s + x\ddot{y}_s + \cdots = \frac{x\dot{F}_s + \frac{x^2}{2}\ddot{F}_s + \cdots}{x\dot{G}_s + \frac{x^2}{2}\ddot{G}_s + \cdots} = \frac{\dot{F}_s + \frac{x}{2}\ddot{F}_s + \cdots}{\dot{G}_s + \frac{x}{2}\ddot{G}_s + \cdots},$$

with $\dot{y}_s = \dot{y}(\mathbf{x}_s)$ et $\ddot{y}_s = \ddot{y}(\mathbf{x}_s)$. Computing $\dot{F}_s$ requires the computation of compound derivatives:

$$\dot{F} = \frac{\partial f}{\partial x} + \dot{y}\frac{\partial f}{\partial y},$$

$$\ddot{F} = \frac{\partial \dot{F}}{\partial x} + \dot{y}\frac{\partial \dot{F}}{\partial y} + \ddot{y}\frac{\partial \dot{F}}{\partial \dot{y}}.$$

We do the same with G. We wish to compute the series expansion of the asymptotic curve at the singular point, that is, an equation of the form $y = y_s + m(x - x_s) + p(x - x_s)^2/2$, with $m = \dot{y}_s = \dot{y}(x_s)$ and $p = \ddot{y}_s = \ddot{y}(x_s)$. To order 0, we must solve the second-order equation:

$$m = \frac{f_x + m\ f_y}{g_x + m\ g_y}, \tag{A.6}$$

to find m. Once m is known, we can infer p, which is solution to the following equation:

$$\ddot{F}_s = m\ddot{G}_s + 2p\dot{G}_s. \tag{A.7}$$

♣ **Example**. – Let us consider the differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{x + 3xy + 3(1-y)y}{3x(2x+3y)}.$$

We would like to determine how the integral curves behave close the origin point (which is singular). We find:

$$\dot{F}(0) = 1 + 3m \ \text{ et } \ \dot{G}(0) = 0,$$

and the solution to (A.6) is $m = -1/3$. To order 1, we get

$$\ddot{F}(0) = -\frac{8}{3}p \ \text{ et } \ \ddot{G}(0) = 6,$$

and the solution to ((A.7) is $p = 2/9$. The equation of the asymptotic curve is then:

$$y = -\frac{1}{3}x\left(1 - \frac{2}{3}x + \cdots\right).$$

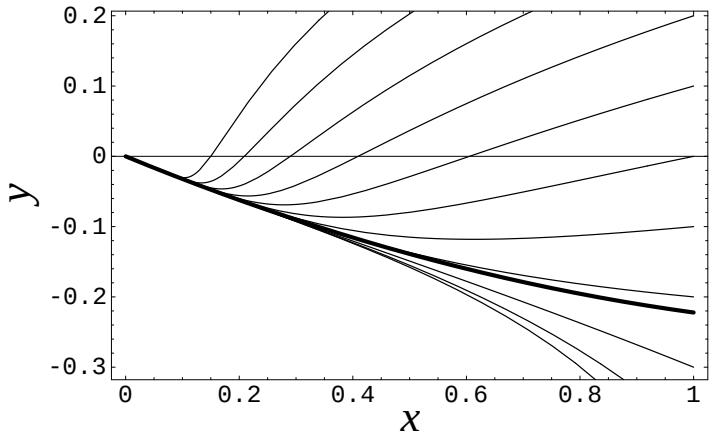


Figure A.7. *Convergence of integral curves towards the asymptotic curve.*

A.3.2 Analytical calculation

The separatrix is a curve, which is solution to the differential equation (A.1) for all Lie groups admitted by the same equation. An implicit curve equation $\phi(x,\ y) = 0$ is invariant under a Lie group $X = \xi\partial_x + \eta\partial_y$ if $X\phi = 0$. Using the condition

$$X\phi = \xi\partial_x\phi + \eta\partial_y\phi = 0,$$

we deduce that ϕ is also solution to the following first-order differential equation

$$y' = \frac{\eta(x,\ y)}{\xi(x,\ y)}.$$

The equation of the separatrix is then obtained by substituting y' by η/ξ into (A.1) (Dressler, 1983; Bluman, 1990) :

$$\frac{\eta(x,\ y)}{\xi(x,\ y)} = \frac{f(x,\ y)}{g(x,\ y)}.$$

To find the equation of the separatrix, we must then find all groups that leave Equation (A.1) invariant. This is beyond the scope of the lecture notes and we mention this technique just for completeness.

♣ **Example**. – Let us consider the differential equation

$$y' = \frac{y(x-y^2)}{x^2},$$

which is invariant under transformation $x_1 = \lambda x$ and $y_1 = \sqrt{\lambda} y$, whose infinitesimal generator is $X = 2x\partial_x + y\partial_y$. We then deduce $\xi = 2x$ and $\eta = y$. The equation of the separatrix is

$$\frac{y}{2x} = \frac{y(x-y^2)}{x^2},$$

or, equivalently,

$$y^2 = \frac{x}{2}.$$

□

A.4 Singular points located at infinity

In many cases, we have $f(x,\ y) \to \infty$ and $g(x,\ y) \to \infty$ when $x \to \infty$ and $y \to \infty$, which implies that the behavior of $\mathrm{d}y/\mathrm{d}x$ is indefinite there. One way to find the proper limit is to use the dominant-balance technique, i.e. $y \ll x$, $y \sim x$, or $y \gg x$, then integrate the resulting differential equation to check whether the assumption is consistent *a posteriori* or not. In some cases, singular points expelled to infinity may actually represent a single point; a change of variable can usually show that (Lacey *et al.*, 1982). For example, with the following variable change

$$x_1 = \frac{x}{x^2+y^2} \text{ et } y_1 = -\frac{y}{x^2+y^2},$$

then by analyzing behavior at $(0,\ 0)$ in the $(x_1,\ y_1)$ plane, we can determine the behavior of a singular at infinity (note that it is equivalent to making the change $z_1 = 1/z$ with $z = x + \imath y$).

♣ **Example**. – Let us consider the differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{3y(x-2y)}{(1-3x)y - x - x^2},$$

for which we consider one of the following possibilities when $x \to \infty$ and $y \to \infty$:

- $y \ll x$, then we find $\dot{y} \propto 3y/x$, i.e., $y \propto x^3$, which contrasts with the initial assumption;

- $y \sim x$, then we set $y \propto mx$, we deduce $m = 4/3$, which is consistent with our assumption;
- $y \gg x$, then $\dot{y}/y \propto 2y/x$), i.e. $y \propto x^2$, which is consistent with our assumption.

For the first quadrant $(x,\ y) > 0$, $x \to \infty$ and $y \to \infty$, there are two singular points at infinity corresponding to two asymptotic curves $y = 4x/3$ and $y = x^2$. For the second quadrant $x \to -\infty$ and $y \to \infty$, there is one singular point corresponding to the curve $y = x^2$, identical to the asymptotic curve we have found previously. It is in fact the same point and in practice, this means that one path may escape from the first quadrant along the curve $y = x^2$ to return via the second quadrant and following the same curve. To show this, we can make the change of variable

$$u = \frac{1}{x} \text{ and } v = \frac{x^2}{y},$$

such that both singular points of the first problem collapse onto a single point $A(0,\ 1)$. We then conclude:

$$\frac{\mathrm{d}v}{\mathrm{d}u} = \frac{v\ (2 - (5 + 2\ u)\ v)}{3 + u\ (-1 + v) + u^2\ v}.$$

Point A is not a singular point of this equation; therefore, for A and for points in the near vicinity of A, only one curve passes.

A.5 Singular points with horizontal/vertical tangent

We can find differential equations with singularities satisfying $m = 0$ or $m = \pm\infty$ and in this case, the behavior is deduced by approximation and integration of the solution (argument like the one used in the *dominant balance* technique).

♣ **Example**. – Let us consider the differential equation:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{8 - 3x}{x(4 - x) - 2}y, \tag{A.8}$$

for which we note that the denominator vanishes at $A_-\ (2-\sqrt{2},\ 0)$ and $A_+\ (2+\sqrt{2},\ 0)$, which are two singular points (nodes). The numerator vanishes at $A_0\ (8/\ 3,\ 0)$, which corresponds to an extremum in the integral curve. The solution has the following behavior:

		A_-		A_0		A_+	
numerator	+	|	+	|	−	|	−
denominator	−	|	+	|	+	|	−
function	−	|	+	|	−	|	+

The behavior around the nodes is then given by:

- for A_-, we get:

$$\frac{\mathrm{d}y}{\mathrm{d}x} \approx n\frac{y}{x - x_{A-}},$$

with $x_{A-} = 2 - \sqrt{2}$ and $n = (8 - 3x_{A-})/(x_{A+} - x_{A-}) = 3/2 + 1/\sqrt{2} \approx 2,20 > 1$. After integration, we find: $y = c(x - x_{A-})^n$, with c an integration constante, thus at point A_-, the curve admits a horizontal tangent.

- for A_+, we have:

$$\frac{\mathrm{d}y}{\mathrm{d}x} \approx n' \frac{y}{x - x_{A+}},$$

avec $n' = (3x_{A+} - 8)/(x_{A+} - x_{A-}) = -3/2 + 1/\sqrt{2} \approx 0,79 < 1$. After integration, we get: $y = c|x_{A+} - x|^{n'}$, thus at point A_+, the curve admits a vertical tangent.

Note that in this case, there is an analytic solution of the form:

$$y(x) = c|2 - ax + x^2|^{3/2} \exp\left[-\sqrt{2}\,\mathrm{arctanh}\frac{x-2}{\sqrt{2}}\right].$$

The result (numerical integration) is reported on the following figure. Note that the vertical tangent at point A_+ is not very apparent because the interval over which the derivative is very large is narrow. □

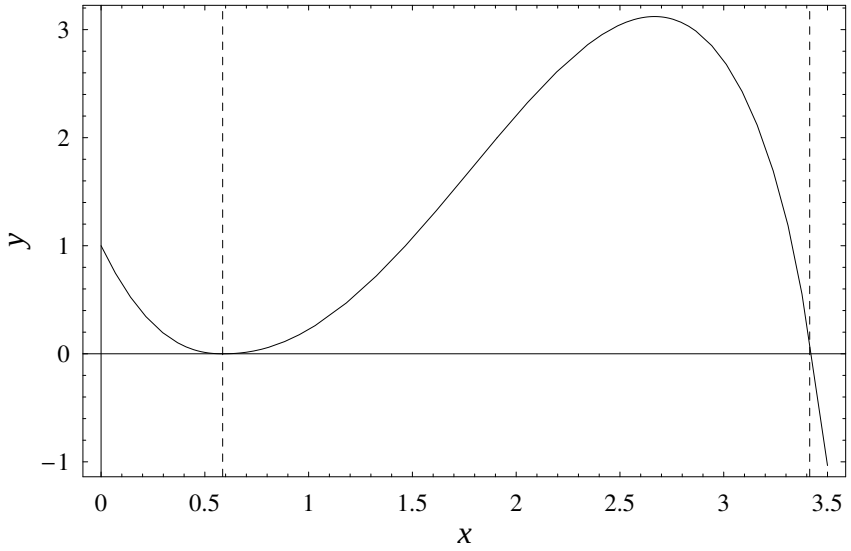


Figure A.8. *Solution to equation (A.8).*

♣ **Example**. – Let us consider the differential equation:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{6y(2y - x)}{2q^2 + 6yx + x}.$$

It is singular at point O and A $(0, \ -1/2)$ At O, we find that $m = 0$. To work out the asymptotic-curve equation, we try to approximate the solution. We assume $x \ll y$ when $y \to 0$. This therefore gives

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{6y(2y - x)}{2q^2 + 6yx + x} \approx \frac{12y^2}{x},$$

whose first integral is the family of curves

$$x = K \exp\left(-\frac{1}{12y}\right),$$

with K an integration constant.

B Approximation

B.1 Definitions

We are concerned with the scalar function (u, ϵ), which takes value x on an interval D and depends on a parameter ϵ over an interval I: $0 < \epsilon < \epsilon_1$.

The notation

$$u(x, \epsilon) = O(v(x, \epsilon)) \text{ sur } I \tag{B.1}$$

means that for all x over D, there exists a number $k(x)$ such that

$$u(x, \epsilon) \leq k(x)|v(x, \epsilon)|, \text{ pour tout } \epsilon \in I.$$

Similarly we have

$$u(x, \epsilon) = O(v(x, \epsilon)) \text{ quand } \epsilon \to 0,$$

for any x of D, when there exists a positive real $k(x)$ and a neighborhood N of $\epsilon = 0$ such that $u(x, \epsilon) = k(x)|v(x, \epsilon)|$ for any ϵ over I. The relation (B.1) is said to be *uniformly valid* when k does not depend on x.

♣ **Example**. – The relation

$$\frac{1}{x + \epsilon} = O(1),$$

is correct because if we take $k(x) = 1/x$ and since $(x + \epsilon)^{-1} < x^{-1}$, the relation $u(x,\ \epsilon) < k(x) \times 1$ holds. However, this relation is not uniformly valide since there is no constant k such that $u(x,\ \epsilon) < k \times 1$. □

The notation $u = O(v)$ does not mean that u and v must have the same order of magnitude. This notation implies tat u is bounded. If u and v have the same order of magnitude, then the limit u/v (when $\epsilon \to 0$) exists and is finite; we have $u = O(v)$ and $v = O(u)$. We can make use of the following notation $u = \text{ord}(v)$ (this notation is, however, not universally accepted).

$$u(x, \epsilon) = o(v(x, \epsilon)) \text{ quand } \epsilon \to 0 \tag{B.2}$$

means that for any x over D and for any constant $k > 0$, there is an interval N_ϵ $(0 < \epsilon < \epsilon_1(x, k))$ such that

$$u(x, \epsilon) \leq k|v(x, \epsilon)|, \text{ for any } \epsilon \in N_\epsilon.$$

u becomes arbitrarily small compared to $|v|$. We also note that $u \ll v$. The relation (B.2) is said to be uniformly valid if ϵ_1 depends on k, but not on x.

A sequence of functions $\phi_n(\epsilon)$ is called *asymptotic series* if

$$\phi_{n+1}(\epsilon) = o(\phi_n(\epsilon)) \text{ quand } \epsilon \to 0.$$

♣ **Example**. – The sequence $\phi_n(\epsilon) = \epsilon^{n-1}$ with $n = 1, 2, \ldots$ is an asymptotic series. □

For a function (u, ϵ), its asymptotic expansion to order N means that when $\epsilon \to 0$, we are able to build a function $\sum_{n=1}^{N} u_n(x)\phi_n(\epsilon)$ such that

$$u(x,\epsilon) - \sum_{n=1}^{N} u_n(x)\phi_n(\epsilon) = o(\phi_M) \text{ quand } \epsilon \to 0,$$

for $M = 1, 2, \ldots, N$ and where $\phi_n(\epsilon)$ is a sequence of functions.

When the function u is known and if the asymptotic series $\phi_n(\epsilon)$ is specified, we can easily construct the sequence of functions $u_n(x)$

$$\begin{aligned}
u_1(x) &= \lim_{\epsilon\to 0} \frac{u(x,\epsilon)}{\phi_1(\epsilon)}, \\
u_2(x) &= \lim_{\epsilon\to 0} \frac{u(x,\epsilon) - u_1(x)\phi_1(\epsilon)}{\phi_2(\epsilon)}, \\
u_k(x) &= \lim_{\epsilon\to 0} \frac{u(x,\epsilon) - \sum_{n=1}^{k-1} u_n(x)\phi_n(\epsilon)}{\phi_k(\epsilon)}.
\end{aligned}$$

B.2 Regular perturbation techniques

Let us consider a projectile launched vertically from the Earth surface (sphere of radius R). Its equation of motion is

$$\frac{\mathrm{d}^2 z}{\mathrm{d}\tau^2} = -\frac{gR^2}{(z+R)^2}, \tag{B.3}$$

with g gravitational acceleration, $\tau > 0$ the time, and with the initial conditions $z(0) = 0$ and $\dot{z}(0) = v_0$. This problem is formally equivalent to

$$\frac{\mathrm{d}^2 y}{\mathrm{d}t^2} = -\frac{1}{(\epsilon y + 1)^2}, \tag{B.4}$$

where $\epsilon = v_0^2/Rg$ is a small parameter ($\epsilon \ll 1$), $t = \tau/T_*$, $z = y/L_*$, $T_* = v_0/g$, and $L_* = v_0^2/g$. We seek an asymptotic approximation to

$$y = y_0 + \epsilon y_1 + \epsilon^2 y_2 + \ldots$$

Substituting this expression into (B.3), we have

$$1 + \left(y_1{}''(t) - 2y_0(t)\right)\epsilon + \left(3y_0(t)^2 - 2y_1(t) + y_2{}''(t)\right)\epsilon^2 + + \ldots = -y_0{}''(x).$$

To order ϵ^0, we have

$$y_0'' + 1 = 0,$$

whose solution is

$$y_0 = t - \frac{1}{2}t^2.$$

To order ϵ, we get

$$y_1'' = 2y_0 = 2t - t^2,$$

whose solution is

$$y_1 = \frac{1}{12}\left(4t^3 - t^4\right).$$

To order ϵ^2, we get

$$y_2'' = 2y_1 - 3y_0^3 = \frac{1}{12}t^2\left(7t^2 - 28t + 36\right),$$

whose solution is

$$y_2 = \frac{1}{360}\left(-7t^6 + 42t^5 - 90t^4\right).$$

The second-order approximation is then

$$y \sim t - \frac{1}{2}t^2 + \frac{\epsilon}{12}\left(4t^3 - t^4\right) + \frac{\epsilon^2}{360}\left(-7t^6 + 42t^5 - 90t^4\right).$$

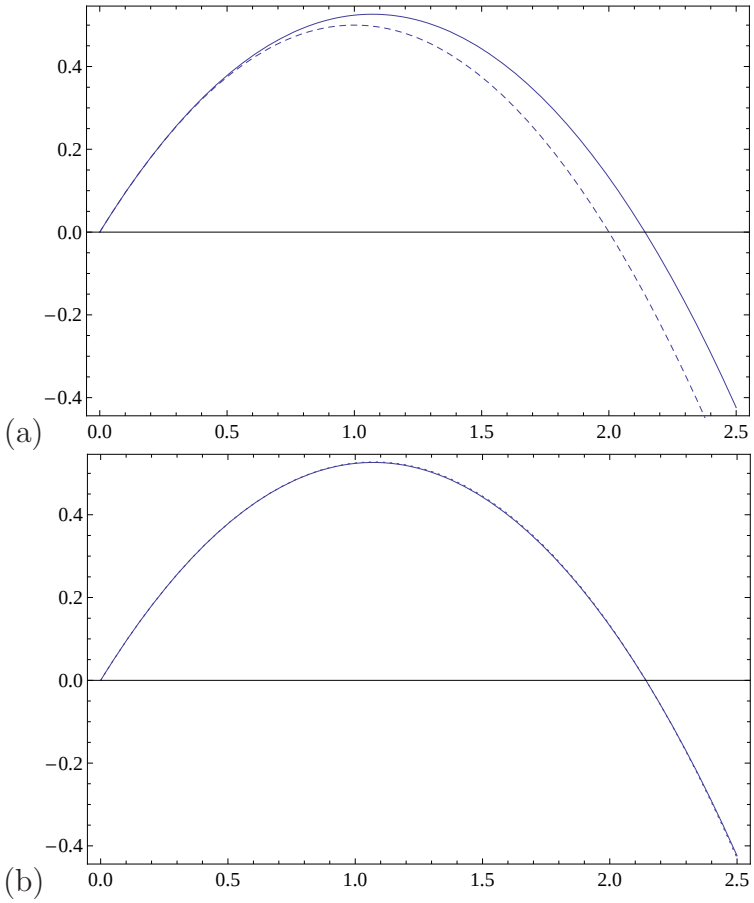


Figure B.1. *Solution to Equation (B.4); dashed line: numerical solution, dashed line (a) solution to order ϵ^0; dotted line (b) solution to order ϵ. Computation done for $\epsilon = 0.1$.*

B.3 Non regular techniques

Let us consider the following differential equation $[0, 1]$

$$(x + \epsilon f)f' + f = 1,$$

with boundary condition $f[1] = 2$. Note that at $x = -\epsilon f$, the differential problem becomes singular. We use the following expansion

$$f = f_0 + \epsilon f_1 + \epsilon^2 f_2 + \dots$$

After substitution, we find that to order 0, we have

$$-1 + f_0 + x f_0' = 0,$$

subject to the condition $f_0(1) = 2$, which yields: $f_0(x) = (1 + x)/x$. To order 1, we find

$$f_1 + x f_1' = \frac{1 + x}{x^3},$$

subject to the condition $f_1(1) = 0$, which gives: $f_1(x) = (3x^2 - 2x - 1)/(2x^3)$. To order 1, we get

$$9 + \frac{4}{x} + 2x^4 f_2 + 2x^5 f_2' = x(2 + 3x),$$

subject to the condition $f_2(1) = 0$, which gives: $f_2(x) = (1 + 3x - x^2 - 3x^3)/(2x^5)$. We finally obtain

$$f(x) = \frac{1 + x}{x} + \epsilon \frac{3x^2 - 2x - 1}{2x^3} + \epsilon^2 \frac{1 + 3x - x^2 - 3x^3}{2x^5} + o(\epsilon^2). \tag{B.5}$$

This approximation is asymptotic when x is fixed and $\epsilon \to 0$, but is no longer asymptotic when $x = \mathcal{O}(\epsilon^{1/2})$. Indeed, the second term in the approximation, which should be $O(\epsilon)$ becomes $O(1)$. To address this issue more carefully, we can compare the approximate solution with the exact solution. The latter could be easily obtained by making use of $(x+\epsilon f)f' + f = 1$, which is equivalent to $\epsilon \frac{1}{2}(f^2)' + (xf)' = 1$. It is easy to integrate this equation: $\epsilon \frac{1}{2} f^2 + xf = x + c$, with c a constant. We find

$$f(x) = \frac{\sqrt{x^2 + 2\epsilon x + 4\epsilon^2 + 2\epsilon} - x}{\epsilon}.$$

The solution comparison reported on Fig. B.2 shows that the approximate solution diverges when approaching the origin point, whereas the exact solution does not do so. The approximate solution is singular at 0.

We can get around this delicate issue using two different methods.

B.3.1 Stretched-coordinate method

We make use of the two following expansions

$$f(x, \epsilon) = f_0(s) + \epsilon f_1(s) + \dots,$$
$$x(s, \epsilon) = s + \epsilon x_1(s) + \dots,$$

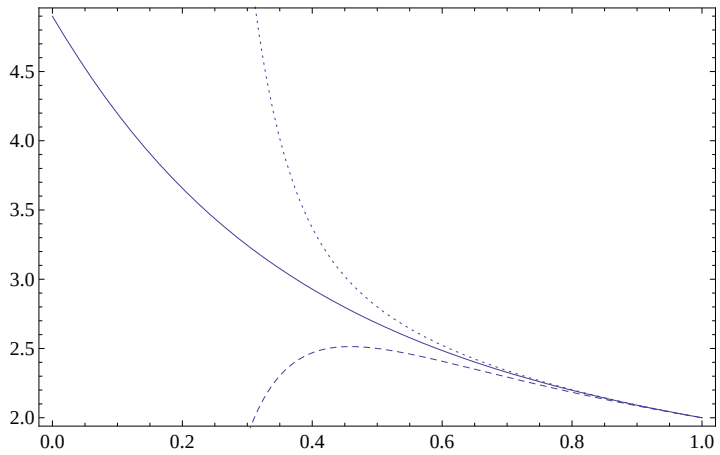


Figure B.2. *Solution to Equation* $(x + \epsilon f)f' + f = 1$. *Solid line: exact solution. Dashed line: approximate solution to ordre 1. Dotted line: approximate solution to ordre 2. Computation done for* $\epsilon = 0.1$.

with f_0 that must be uniformly asymptotic over $[0, 1]$. The differential operator

$$\frac{\mathrm{d}}{\mathrm{d}x} = \frac{\mathrm{d}s}{\mathrm{d}x}\frac{\mathrm{d}}{\mathrm{d}s} = \frac{1}{\mathrm{d}x/\mathrm{d}s}\frac{\mathrm{d}}{\mathrm{d}s} = (1 - \epsilon x_1' + \epsilon^2(x_1'^2 - x_2')\ldots)\frac{\mathrm{d}}{\mathrm{d}s}.$$

To order 0, we find

$$sf_0' + f_0 = 1,$$

with $f_0(1) = 2$, which yields: $f_0(s) = (1 + s)/s$. To order ϵ, we get:

$$f_1 + sf_1' + (f_0 + x_1)f_0' - sf_0'x_1' = 0,$$

subject to boundary conditions

$$\begin{aligned} 2 &= f_0(s) + \epsilon f_1(s) + \ldots \\ &= f_0(1 - \epsilon x_1(s)) + \epsilon f_1(1 - \epsilon x_1(s)) + \ldots \\ &= f_0(1) + \epsilon\left(-x_1(1)f_0'(1) + f_1(1)\right) + O(\epsilon^2), \\ 1 &= s + \epsilon x_1(s) + \ldots. \end{aligned}$$

We then derive $f_1(1) = -x_1(1)$. Substituting f_0 into this expression leads to

$$f_1 + sf_1' = \frac{x_1}{s^2} - \frac{x_1'}{s} + \frac{1}{s^2} + \frac{1}{s^3}, \tag{B.6}$$

One way of proceeding is to select x_1 such that the governing equation of f_1 is homogeneous. Solving

$$\frac{x_1}{s^2} - \frac{x_1'}{s} + \frac{1}{s^2} + \frac{1}{s^3} = 0$$

gives us

$$x_1 = \frac{3s^2 - 2s - 1}{2s}.$$

Integrating Equation (B.6) provides $f_1 = 0$. When we invert $x = s + \epsilon x_1(s)$, we find

$$s = \frac{-x + \epsilon + \sqrt{x^2 + 4\epsilon^2 - 2(x + 1)\epsilon}}{3\epsilon - 2},$$

which, once substituted into the equation of f_0, gives us

$$f \sim \frac{x + \sqrt{x^2 + 4\epsilon^2 - 2(x+1)\epsilon}}{\epsilon},$$

which is, in fact, identical to the analytical solution.

There is no unique choice of x_1. We could have proceeded differently. For instance, we can integrate Equation (B.6) easily

$$(sf_1)' = \left(-\frac{x_1}{s} - \frac{1}{s} - \frac{1}{2s^2}\right)',$$

that is,

$$f_1 = \frac{A}{s} - \frac{x_1}{s^2} - \frac{1}{s^2} - \frac{1}{2s^3},$$

for which the boundary conditions yield $A = 0$. To get rid of the singularities (i.e., terms of order of magnitude larger than $1/x$), we set

$$x_1 = -\frac{1+2s}{2s},$$

which yields $f_1(s) = 3/(2s)$. We can invert $x = s + \epsilon x_1(s)$ and substitute into f_0

$$f(x) \sim \frac{x + \epsilon + \sqrt{x^2 + 2\epsilon x + \epsilon(\epsilon+2)} + 2}{x + \epsilon + \sqrt{x^2 + 2\epsilon x + \epsilon(\epsilon+2)}} = -\frac{x}{\epsilon} + \frac{1}{\epsilon}\sqrt{x^2 + 2\epsilon x + \epsilon(\epsilon+2)}.$$

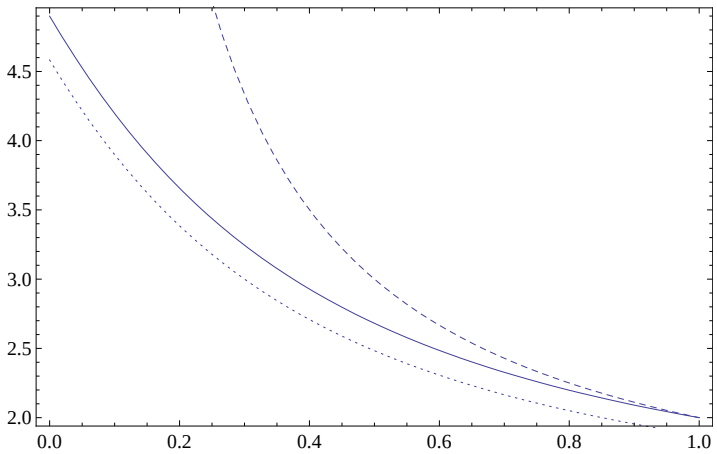


Figure B.3. *Solution to Equation $(x + \epsilon f)f' + f = 1$. Solid line: exact solution. Dashed line: approximate solution to ordre 1. Dotted line: approximate solution to ordre 2. Computation done for $\epsilon = 0.1$.*

B.3.2 Matched perturbation technique

The approximate solution is no longer valid for $x = \mathcal{O}(\epsilon^{1/2})$ where $f = \mathcal{O}(\epsilon^{-1/2})$. This leads us to make use of new variables, which will be used for the *inner solution*, i.e., the solution close to the singular point:

$$x = \epsilon^{1/2}\xi,$$
$$f_i(x) = \epsilon^{-1/2}F_0(\xi) + F_1(\xi) + \epsilon^{1/2}F_2(\xi) + \dots.$$

Since $\mathrm{d}/\mathrm{d}x = \epsilon^{-1/2}\mathrm{d}/\mathrm{d}\xi$, we deduce that to order $\epsilon^{-1/2}$, we have

$$F_0 + \xi F0' + F_0 F_0' = 0,$$

whose general solution is

$$F_0 = \sqrt{\xi^2 + A_0} - \xi,$$

with A_0 a constant of integration. To order ϵ^0, we get

$$-1 + F_1 + F_1 F_0' + \xi F_1' + F_0 F_1' = 0,$$

whose general solution is

$$F_1 = \frac{A_1 + \xi}{\sqrt{\xi^2 + A_0}},$$

with A_1 a constant of integration To order $\epsilon^{1/2}$, we have

$$F_2 + F_2 F_0' + F_1 F_1' + \xi F_2' + F_0 F_2' = 0,$$

whose general solution is

$$F_2 = \frac{-A_1^2 - 2\xi A_1 + 2A_2\xi^2 + A_0 + 2A_0 A_2}{2\left(\xi^2 + A_0\right)^{3/2}},$$

with A_2 a constant of integration.

We now have to determine the constants of integration. Take a closer look at the *outer solution*

$$f(x) = \frac{1+x}{x} + \epsilon\frac{3x^2 - 2x - 1}{2x^3} + O(\epsilon^2).$$

The change of variable $x = \epsilon^{1/2}\xi$ shows that

- to order $\epsilon^{-1/2}$ and in the limit $\xi \to \infty$, we have $f_{\epsilon^{-1/2}} \sim \xi^{-1}$ while $F_0 \sim A_0/(2\xi)$ hence $A_0 = 2$;
- to order ϵ^0 and in the limit $\xi \to \infty$, we have $f_{\epsilon^0} \sim 1$ while $F_1 \sim 1$ hence we set $A_1 = 0$;
- to order $\epsilon^{1/2}$ and in the limit $\xi \to \infty$, we have $f_{\epsilon^{1/2}} \sim 3/(2\xi)$ while $F_2 \sim A_2/(\xi)$ hence $A_2 = 3/2$.

The inner solution is

$$f_i(\xi) = \epsilon^{-1/2}\left(\sqrt{\xi^2 + 2} - \xi\right) + \frac{\xi}{\sqrt{\xi^2 + A_0}} + \epsilon^{1/2}\frac{3\xi^2 + 8}{2\left(\xi^2 + 2\right)^{3/2}},$$

while the outer solution f_e is given by (B.5). The composite solution is

$$f = f_i + f_e - f_r,$$

with f_r the matching value between f_i and f_e

$$f_r = \lim_{\xi\to\infty} f_i = \epsilon^{-1/2}\xi^{-1} + 1 = \frac{1}{x} + 1.$$

We note that $f_e = f_r$. The solution is plotted on Fig. B.4. Note that reasonably good agreement is obtained.

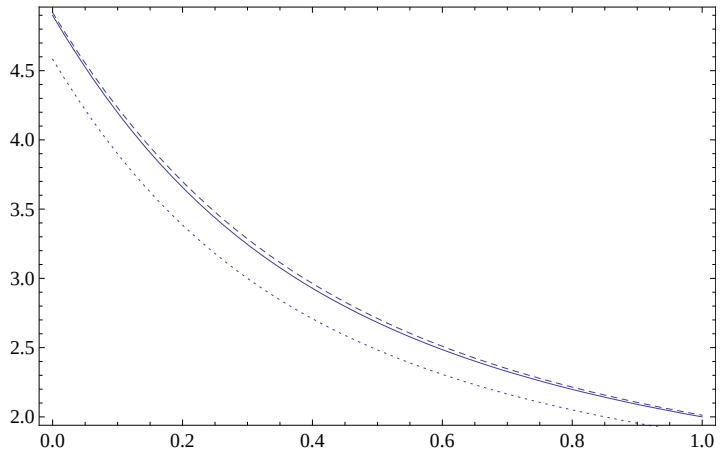


Figure B.4. *Solution to Equation* $(x + \epsilon f)f' + f = 1$. *Solid line: exact solution. Dashed line: approximate solution to ordre* $\epsilon^{1/2}$. *Dotted line: approximate solution to ordre 1 using the stretched coordinate method. Computation done for* $\epsilon = 0.1$.

C Wavelet decomposition

C.1 Introduction to wavelets

C.1.1 Properties

In order to improve the resolution of methods such as the collocation method, we have to use more elaborate or efficient polynomials. An alternative to Legendre polynomial is to use wavelet decomposition. Wavelets are generated from dilations and translations of a special function, called the mother wavelet ψ.

Let us consider a wavelet function ψ with a support I_ψ; this wavelet function, usually referred to as the *mother wavelet*, is associated with a *scaling function* ϕ with a support $I_\phi = [0,\ B]$. We can define a family of orthogonal (in a sense that is specified below) functions:

$$\psi_{ij}(x) = 2^{i/2}\psi(2^i x - j) \text{ and } \phi_{ij}(x) = 2^{i/2}\phi(2^i x - j).$$

These functions enjoy helpful properties:

- pairwise orthogonality: $< \psi_{ij},\ \psi_{kl} >= \delta_{ik}\delta_{jl}$ and $< \phi_{ij},\ \phi_{ik} >= \delta_{jk}$
- orthogonality $< \phi_{ij},\ \psi_{kl} >= 0$ for $k \geq i$
- 0-moments: $\int_{\mathbb{R}} \psi(x)\mathrm{d}x = 0$ and $\int_{\mathbb{R}} \phi(x)\mathrm{d}x = 1$
- normality: $\int_{\mathbb{R}} \psi^2(x)\mathrm{d}x = 1$ and $\int_{\mathbb{R}} \phi^2(x)\mathrm{d}x = 1$
- scaling properties: there is a (usually finite) set of coefficients h_k such that:

$$\phi(x) = \sum_{k\in\mathbb{Z}} h_k \sqrt{2}\phi(2x - k). \tag{C.1}$$

The last property is important because it allows the *multiresolution analysis* (Vidakovic, 1999; Mallat, 1998). The coefficients h_k are called the *wavelet filter*; they act as a low-pass averaging filter. We introduce n_h the length of nonzero wavelet coefficients h_k. The support of ϕ is directly related no n_h: $\mathrm{supp}\phi = [0,\ n_h - 1]$; the support of the mother wavelet is $[-n_h/2,\ n_h/2]$. A key point is that it is possible to relate $\phi(x)$ to $\phi(2x - k)$, i.e. by averaging what happens on a fine scale, we can obtain the trend. Another aspect is that this relationship is a convolution; therefore, if we work in the Fourier domain, we obtain the relationship:

$$\hat{\phi}(\omega) = m_0(\omega)\hat{\phi}\left(\frac{\omega}{2}\right),$$

where $\hat{\phi}(\omega)$ is the Fourier transform of ϕ and m_0 is the discrete Fourier transform of (h_k).

In many applications, one tries to find sparse representations of functions, i.e., approximations with a few non-zero coefficients. It is then of great importance to select wavelet bases that efficiently approximate particular classes of functions. Wavelet design is optimized to produce a maximum of coefficient $< f,\ \psi_{ij} >$ that are zero or close to zero. An important property is the number p of vanishing moments of the mother wavelet ψ. The i-moment is: $M_i =< x^i,\ \psi(x) >= \int_{\mathbb{R}} \psi(x) x^i \mathrm{d}x$. If ψ has p vanishing moments, then ψ is orthogonal to any polynomial of degree $p-1$. If a function is is locally $\mathcal{C}^k$, then over a small interval it is approximated by a polynomial of degree k. Therefore, for $k < p$, we have:

$$< f,\ \psi_{ij} >= 2^{i/2} \int_R f(x)\psi(2^i x - j)\mathrm{d}x = 2^{i/2} \int_R f(x + 2^{-i}j)\psi(2^i x)\mathrm{d}x$$

$$< f,\ \psi_{ij} >= 2^{i/2} \sum_{i=0}^{k} \int_R f^{(n)}(2^{-i}j) x^n \psi(2^i x)\mathrm{d}x + \mathcal{O}(x^p) = 2^{-i/2} \sum_{i=0}^{k} f^{(n)}(2^{-i}j) M_n + \mathcal{O}(x^p)$$

Finally we deduce that: $< f,\ \psi_{ij} >= \mathcal{O}(x^p)$. In other words, if the mother wavelet has p vanishing moments, then any polynomial of degree $p-1$ is fully reproduced by the scaling function. It can be shown that the following properties are equivalent:

1. the mother wavelet has p vanishing moments
2. $\hat{\psi}(\omega)$ and its first $p-1$ derivatives are zero at $\omega = 0$
3. $\mathrm{supp}\psi \geq 2p-1$
4. for any $0 \leq k < p$, $q_k(t) = \sum_{n\in\mathbb{Z}} n^k \phi(t-n)$ is a polynomial of degree k.

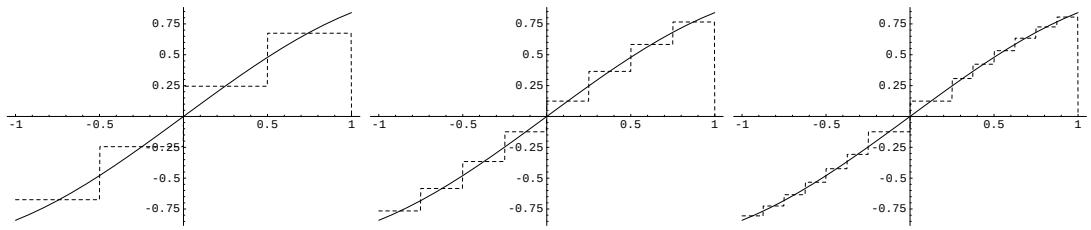
The last property shows how is possible to exactly represent polynomials $\{1,\ t,\ t^2,\ \ldots\}$ by wavelets. Daubechies wavelets are optimal because they have a minimum size support (according to property (3), $\mathrm{supp}\psi = 2p-1$) for p vanishing moments[1]: $\mathrm{supp}\phi[\mathrm{D}p] = [0,\ 2p-1]$ and $\mathrm{supp}\psi[\mathrm{D}p] = [-p+1,\ p]$.

When choosing a particular wavelet, we have to find a trade-off between the number of vanishing moments and the support width. Indeed, the wider the support of ψ compared to the support of a function f, the more numerous the shift index required to span $\mathrm{supp}f$. Furthermore if f has few isolated singularities and is very regular, it is more convenient to select a wavelet with a large support and many vanishing moments to provide a sparse representation of f. Otherwise, it may be better to decrease the support width of ψ to single out singularities and avoid high amplitude coefficients.

C.1.2 Multiresolution analysis

Multiresolution analysis is a technique that provides approximations of signals at various resolution levels by using orthogonal projections onto different spaces. These projections can be computed by using projector operators in a given function space, but it can also be shown that multiresolution approximations are entirely characterized by the wavelet filter, which can be interpreted as a operator controlling the loss of information across different levels (see

[1]This property is easy generalized to any wavelet whose scaling function has support $[N_1,\ N_2]$; in that case, we have: $\mathrm{supp}\psi = [(N_1 - N_2 + 1)/2,\ (N_2 - N_1 + 1)/2]$.



$i = 0$ $i = 1$ $i = 2$

Figure C.1. *Decomposition of $f(x) = \sin x$ defined over $[-1,\ 1]$ into Haar wavelets (up to $i = 2$).*

§ C.1.3). The idea is to consider that the approximation of a function f at a resolution level 2^i is equivalent to a local average of f over neighborhoods of size 2^i.

It is possible to decompose any function of $f(x)$ of $\mathcal{L}^2[a,\ b]$:

$$f(x) = \sum_{i=-\infty}^{\infty} \sum_{j \in J} \beta_{ij} \psi_{ij}(x)$$

where J is a set of j-index for which $0 \leq 2^i x - j \leq B$ for a given scale i and $\beta_{ij} = \int_{\mathbb{R}} \psi_{ij}(x) f(x) \mathrm{d}x$. The problem is that the summation is made over infinite sets. It can be shown that:

$$f(x) = \sum_{j \in J_0} \alpha_{i_0 j} \phi_{i_0 j}(x) + \sum_{i \geq i_0}^{\infty} \sum_{j \in J} \beta_{ij} \psi_{ij}(x)$$

The first term on the right-hand side of the equation is called the *trend*, i.e., it represents the mean or filtered (low-pass filter) behavior of the function f at the scale i_0. The second term represents the deviation from this trend; the summation is made over different scale i. J_i denotes the set of j-index needed to describe f at scale i. Since $\alpha_{ij} =< f,\ \phi_{ij} >$, we have:

$$\alpha_{ij} = \int_{\mathbb{R}} f(x) \phi_{ij}(x) \mathrm{d}x = 2^{i/2} \int_a^b f(x) \phi(2^i x - j) \mathrm{d}x = 2^{-i/2} \int_{2^i a - j}^{2^i b - j} f\left(\frac{\xi + j}{2^{-i}} \right) \phi(\xi) \mathrm{d}\xi$$

For this integral not to be zero, the bounds must verify: $2^i a - j < B$ and $2^i b - j > 0$, i.e.:

$$J_i : \quad 2^i a - B < j < 2^i b.$$

For instance, let us consider the Haar wavelets ($I_\psi = I_\phi = [0,\ 1]$) and a function with a support over $[-1,\ 1]$. At each scale i, we define the set $J_i : \quad 2^i - 1 < j < 2^i$. Thus at scale $i = 0$, $J_0 = \{-1,\ 0\}$, $i = 1$, $J_1 = \{-2,\ -1,\ 0,\ 1\}$, $i = 2$, $J_2 = \{-4,\ -3,\ -2,\ -1,\ 0,\ 1,\ 2,\ 3\}$, etc. For instance for $f(x) = \sin x$, at $i = 1$, one has (see Fig. C.1):

$$\begin{aligned} f(x) = {} & 0.459(\phi_{0,0}(x) - \phi_{0,-1}(x)) - 0.214(\psi_{0,0}(x) - \psi_{0,-1}(x)) \\ & -0.064(\psi_{1,-2}(x) - \psi_{1,1}(x)) - 0.085(\psi_{1,-1}(x) - \psi_{1,0}(x)) + \cdots \end{aligned}$$

C.1.3 Multiresolution analysis for finite samples

In practice, one has a sampled function $f(x_i)$ for $i = 1 \cdots n$ (here we assume that $n = 2^N$) and we want to obtain a wavelet decomposition from this sample. Here we assume that the

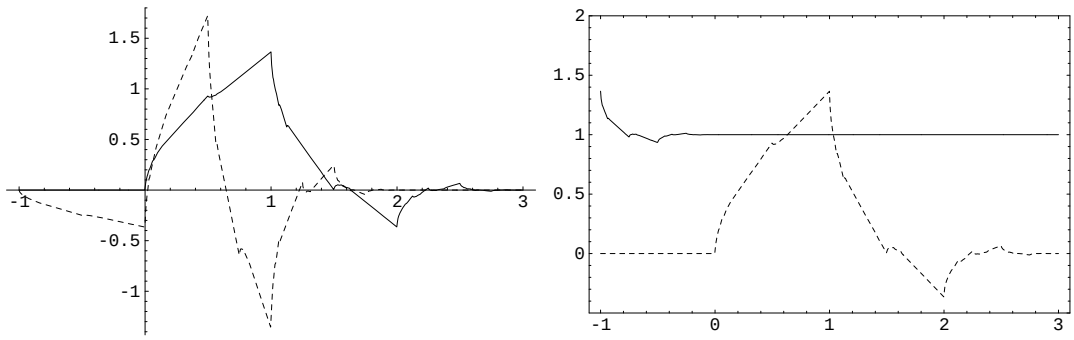


Figure C.2. *Left: scaling function (solid line) and wavelet (dashed line) for the Daubechies wavelet D2. Right: periodized scaling function (solid line) $\psi_{00}^{per}(x)$ compared to the original scaling function ϕ.*

wavelet support is $[0,\ 1]$. Except for some basic wavelets such as the Haar wavelets, the support is usually different from $[0,\ 1]$. It is, however, possible to still work on the interval $[0,\ 1]$ by transforming (periodizing) the wavelets in the following way [see (Mallat, 1998), § 7.5.1; (Vidakovic, 1999), § 5.6]:

$$\psi_{ij}^{per}(x) = \sum_{k\in\mathbb{Z}} 2^{i/2}\psi(2^i(x+k)-j)$$

for the wavelets ψ_{ij} whose support intersects or includes the interval $[0,\ 1]$. When the support of the wavelet at scale i and for shift index j lies within $[0,\ 1]$, then the function ψ_{ij} is preserved.

We report a typical example of periodization for the for the Daubechies wavelet D2 in Fig. C.2. Therefore, in the following, we can consider without loss of generality that the support of the wavelets is $[0,\ 1]$; we no longer use the superscript *per* for the sake of simplicity. Note that this choice entails some inconvenients: near the boundaries, periodic wavelets have a poor behavior because they induce high amplitude coefficients in the neighborhood of the boundaries when the function f is not periodic. It is possible to alleviate these problems by altering the wavelets coefficients close to the boundaries (by using boundary wavelets), but the computations are made more complicated [see (Mallat, 1998), § 7.5.3].

The discrete wavelet transform map the data $\mathbf{d} = (f(x_i))_{1\leq i\leq n}$ to the wavelet domain $\mathbf{w} = (c_{ij},\ d_{ij})$. The result is a vector of the same size n. There is an $n \times n$ orthogonal (or close to orthogonal) matrix $\mathbf{W}$ such that:

$$\mathbf{d} = \mathbf{W}\cdot\mathbf{w},$$

where $W_{pq} = n^{-1/2}\Psi_q(p/n)$ in which $q = (i,\ j)$ is an appropriate set of scale and translation indices. We have introduced the function family Ψ_k that provides a generic representations of the scaling and wavelet functions involved in the decomposition[2]. Here Ψ has to be defined at each scale. At the coarsest level, the scale $i = 0$, we have $\Psi_0 = \phi_{00}$ and $\Psi_1 = \psi_{00}$; at scale $i = 1$, one has $\Psi_2 = \phi_{10}$ and $\Psi_3 = \psi_{11}$; at any scale $i \geq 1$, one has 2^j coefficients d_{ij}, implying that $\Psi_k = \psi_{ij}$ with $k = 2^i + j$ $(0 \leq j \leq 2^i - 1)$.

Let us take a typical example with the Haar wavelet. Consider the function $f(x) = \sin x$ sampled at $x_i = -1+2/(n-1)i$ with $n = 8$ $(N = 3)$. We have: $\mathbf{d}$ ={−0.841, −0.655, −0.415,

[2] For instance, taking the example above of the decomposition of $\sin x$, we set: $\Psi_1 = \phi_{0,-1}$, $\Psi_2 = \phi_{0,0}$, $\Psi_3 = \psi_{0,-1}$, $\Psi_4 = \psi_{2,0}$, etc.

$-0.142,\ 0.142,\ 0.416,\ 0.655,\ 0.841\}$. The functions Ψ are: $\Phi_0 = \phi_{00}$, $\Phi_1 = \psi_{00}$, $\Phi_2 = \psi_{10}$, $\Phi_3 = \psi_{11}$, $\Phi_4 = \psi_{20}$, $\Phi_5 = \psi_{21}$, $\Phi_6 = \psi_{22}$, and $\Phi_7 = \psi_{23}$. We deduce the following 8×8 orthogonal matrix:

$$\mathbf{W} = \begin{bmatrix} \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2} & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{2\sqrt{2}} & \frac{-1}{2\sqrt{2}} & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2\sqrt{2}} & \frac{-1}{2\sqrt{2}} & 0 & \frac{1}{2} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2\sqrt{2}} & \frac{-1}{2\sqrt{2}} & 0 & -\frac{1}{2} & 0 & 0 & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & \frac{-1}{2\sqrt{2}} & 0 & -\frac{1}{2} & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

By taking $\mathbf{w} = \mathbf{W}^* \cdot \mathbf{d}$, one deduces: $\mathbf{w} = \{0,\ -1.452,\ -0.469,\ -0.469,\ -0.131,\ -0.193,\ -0.193,\ -0.131\}$. A tricky point: Can we approximate $f(x)$ by $\sum_i w_i \Psi_i$? In other words, is there a direct link between the discrete wavelet coefficients $\mathbf{w} = (c_{00},\ d_{00},\ \ldots)$ and their continuous counterparts $\mathbf{c} = (\alpha_{ij},\ \beta_{ij})$? The response is positive:

$$\boxed{c_{ij} \approx \sqrt{n}\alpha_{ij} \text{ and } d_{ij} \approx \sqrt{n}\beta_{ij}.} \tag{C.2}$$

To show this, two arguments can be used:

• **Inner product**: the inner product in $\mathcal{L}_2(\mathbb{R})$ is defined by: $< f,\ g >= \int_R f(x)g(x)\mathrm{d}x$. We are interested in its relationship with a discrete equivalent $[f,\ g] = \sum_i f(x_i)g(x_i)$ computed at certain interpolation points x_i regularly spaced by a distance δ. If there are n_k interpolation points, then $\delta \sim (\mathrm{supp}f \cap \mathrm{supp}g)/n_k$, implying $[f,\ g] \propto n_k < f,\ g >$ and $||f|| =< f,\ g >^{1/2} \propto n_k^{1/2}[f,\ f]^{-1/2}$. Continuous and discrete relationships are related via a factor $\sqrt{n}$.

• **Behavior at the finest scale**: let us consider the wavelets coefficients at the finest scale. If we have $n = 2^N$ data, the finest scale is at $i = N$. The continuous wavelet coefficient is in the form:

$$\beta_{ij} =< f(x),\ \psi_{ij} >= 2^{i/2} \int_{\mathbb{R}} f(x)\psi(2^i x - j)\mathrm{d}x = 2^{i/2} \int_{\mathbb{R}} f(x + j2^{-i})\psi(2^i x)\mathrm{d}x$$

If $i = N$ is sufficiently high, $\psi(2^i x)$ is highly concentred in a small region near $x = 0$. So to the leading order in 2^{-i}, we have:

$$\beta_{ij} = 2^{-i/2} f\left(\frac{j}{2^i}\right) + \mathcal{O}(2^{-i}) \tag{C.3}$$

The interpretation is quite simple: at the finest scale $i = N$, the coefficient β_{Nj} correspond the signal measured at dyadic points $j2^{-i}$ multiplied by $2^{-i/2} = 2^{-N/2} = \sqrt{n}$. Here, by definition, the discrete coefficients d_{Nj} at the finest scale are $f(x + j2^{-i})$. We finally deduce:

$$\beta_{Nj} \approx \frac{1}{\sqrt{n}} d_{Nj} \tag{C.4}$$

From the discrete wavelet transform, we can deduce the continuous coefficients by dividing by $\sqrt{n}$, here $2\sqrt{2}$ for our example. We find: $\tilde{\alpha}_{00} = 0,\ \tilde{\beta}_{00} = -0.513,\ \tilde{\beta}_{10} = -0.165,\ \tilde{\beta}_{11} = -0.513$, etc. Furthermore we have to take into account that the decomposition was made

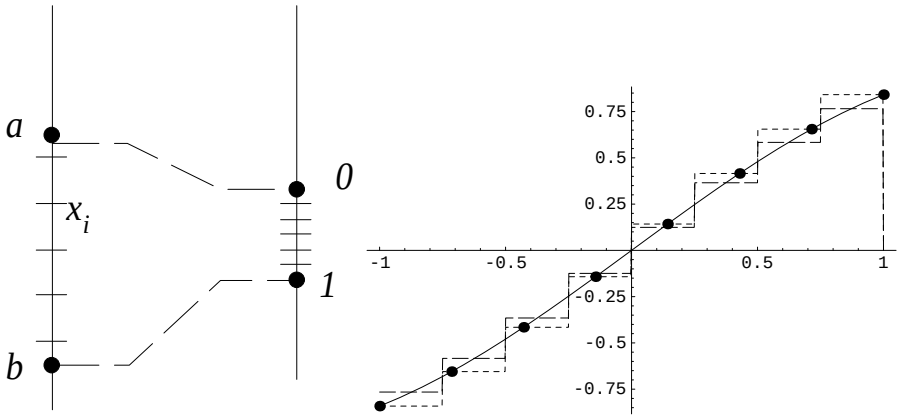


Figure C.3. *Left: in the discrete wavelet decomposition, the interval* $[a,\ b]$ *over which the function is sampled is mapped to* $[0,\ 1]$. *Right: Decomposition of* $f(x_i) = \sin x_i$ *into Haar wavelets, with* $x_i = -1 + 2i/7$ *and* $0 \leq i \leq 7$. *The dots represent the measurements* $f(x_i)$, *the solid line represents the function* sin, *the dashed curve represents the discrete wavelet approximation, and the long-dashed curve is the discrete wavelet decomposition (same curve as in Fig. C.1.*

on a *wrapped* sample $\{i/n,\ f(x_i)\}$, i.e. by transforming the sample for it to lie within $[0,\ 1]$ (see Fig. C.3). At the last stage we have to map $[0,\ 1]$ to the initial interval $[a,\ b]$. Since we have defined $x_i = a + i(b-a)/n$, we have $i/n = (x-a)/(b-a)$. Finally we can express the approximate wavelet transform as:

$$f(x) \approx \sum_i \frac{w_i}{\sqrt{n}} \Psi_i(i/n) = \sum_i \frac{w_i}{\sqrt{n}} \Psi_i \left(\frac{x-a}{b-a} \right)$$

For our example, we arrive at the following approximation:

$$\begin{aligned} \sin(x) \approx & -0.153\psi_{00}(X) - 0.165\,(\psi_{10}(X) + \psi_{11}(X)) \\ & -0.046\,(\psi_{20}(X) + \psi_{23}(X)) - 0.068\,(\psi_{21}(X) + \psi_{22}(X)) \end{aligned}$$

with $X = (x+1)/2$. The resulting curve is plotted in Fig. C.3. Compared to the discrete wavelet decomposition obtained by using the function $\sin(x)$ instead of a sample, there are a small deviations, but they are less than $1/n$ (accuracy of the relationship given in equation C.2).

A major drawback of the Haar decomposition comes of the lack of smoothness. Other bases such as the Daubechies wavelet can be used instead. Figure C.4 shows the comparison between the D2 and Haar wavelets for the function $f(x) = \sin 2x$ sampled at 2^5 points over the interval $[-1,\ 1]$. A better agreement is obtained with the D2 wavelets, but spurious wide fluctuations are introduced at the boundaries.

C.1.4 Cascade algorithm for multiresolution analysis for finite samples

Note that, in practice, the coefficients are computed by a *cascade algorithm* (Mallat, 1998), which is a fast algorithm that requires approximately n operations to compute $\mathbf{w}$ instead of

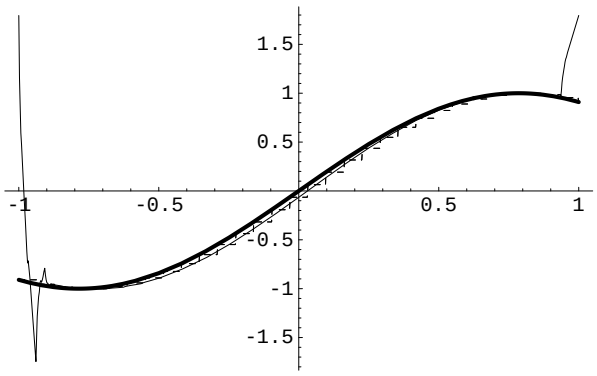


Figure C.4. *Comparison between the D2-wavelet decomposition (solid line) and Haar-wavelet decomposition (dashed line) for the function $f(x) = \sin 2x$ sampled at 2^5 points over the interval $[-1,\ 1]$.*

N^2 if we use matrix operations. If we note $c_{ij} =< f, \phi_{ij} >$ and $d_{ij} =< f, \psi_{ij} >$, we have

$$c_{i-1,j} = \sum_{n\in\mathbb{Z}} h_{n-2j} c_{i,j}$$

$$d_{i-1,j} = \sum_{n\in\mathbb{Z}} g_{n-2j} c_{i,j}$$

where h_i are the wavelet coefficients (see Eq. C.1) and $g_i = (-1)^i h_{1-i}$. We have found that $c_{i+1,j}$ can be computed by taking every convolution of c_{ij} with h. The coefficients pertaining c_{ij} and d_{ij} to resolution level i can be obtained from those at a finer scale $i+1$. If the procedure is repeated, we can recursively build the sequences c_{ij} and d_{ij} provided we specify the initial elements corresponding to the finest scale. Using equations (C.3–C.4), we can take at the finest scale: $d_{ij} = f(j/2^i)$, i.e. the function sampled at the dyadic points $x_i = j/2^i$. A striking point in this cascade algorithm is that we only need h_i and g_i and we do not need compute the inner product $< f, \phi_{ij} >$ and $< f, \psi_{ij} >$.

In practice, different variants can be used:

- let us consider that the number of data is $n = 2^N$ (otherwise we can pad with zeros to obtain an appropriate number). We can repeat the recursive procedure from the finest to the coarsest level. We can consider than the coarsest level is $i = 0$ (described by c_{00} and d_{00}) and the finest $i = N$ (described by c_{Nk} and d_{Nk} for $0 \le k \le N-1$). An alternative is to iterate the decomposition as many times as possible and stop when the length of the trend $c_{i_0,j}$ becomes either odd or smaller than the filter length n_h; this is, for instance, the procedure used in Mathematica.

- another problem concerns the boundaries. For instance, the boundary points can be obtained using:

$$c_{i-1,0} = \sum_{k=0}^{n_h-1-m} h_k c_{ik} \quad \text{and} \quad c_{i-1,2^{i-1}-1} = \sum_{k=0}^{2^i+n_h-1-3-m} h_{k-2^i+2} c_{ik},$$

  where m is an integer $0 \le m \le n_h - 2$. So we need to pad the original data with with m data in the front and $n_h - 2 - m$ data points at the end. The values of data points that are padded at the boundaries are determined by the choice of the boundary condition (periodic, reflective, etc.).

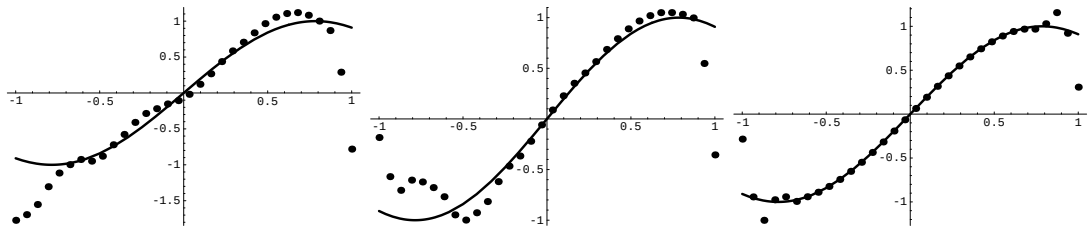


Figure C.5. *Comparison for the D4-wavelet decomposition (dots) between three types of boundary conditions: reflective (on the left), zero (center), periodic (on the right) for the function* $f(x) = \sin 2x$ *sampled at* 2^5 *points over the interval* $[-1,\ 1]$.

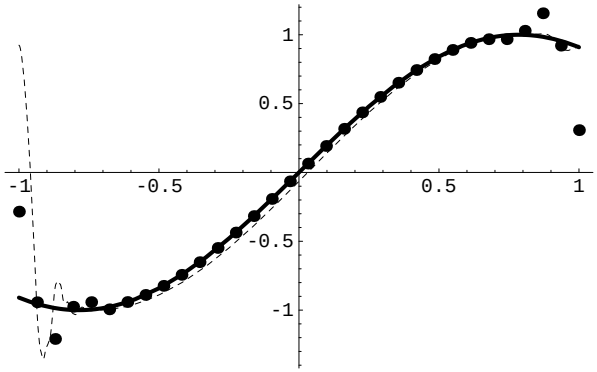


Figure C.6. *Comparison for the function* $f(x) = \sin 2x$ *(solid line) sampled at* 2^5 *points over the interval* $[-1,\ 1]$ *and the D4-wavelet decomposition between the procedure used in Mathematica (dots) and the complete decomposition.*

On Fig. C.5, we have reported three wavelet decomposition for the wavelet basis D4 in the case where we sample the function $f(x) = \sin x$ over the interval $[0,\ 1]$ (here $n = 2^5 data$. Here the better agreement is obtained with periodic boundary conditions. On Fig. C.6, we have plotted the D4-wavelet decomposition for two different assumptions on the coarsest level: in Mathematica, the coarsest level corresponds to the resolution level below which the number of coefficients c_{ij} would become either odd or lower than the filter length. The usual way is to repeat the procedure till there is a single coefficient c_{00}.

Bibliography

Abramovich, F. & Silverman, B.W. 1998 Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85**, 115–129.

Ancey, C. 2005 Solving the Couette inverse problem by using a wavelet-vaguelette decomposition. *J. Rheol.* **49**, 441–460.

Ancey, C., Coussot, P. & Evesque, P. 1996 Examination of the possibility of a fluid-mechanics treatment for dense granular flows. *Mech. Cohesive-Fricct. Mater.* **1**, 385–403.

Ancey, C. & Evesque, P. 2000 Frictional-collisional regime for granular suspension flows down an inclined channel. *Phys. Rev. E* **62**, 8349–8360.

Astarita, G. 1990 The engineering reality of the yield stress. *J. Rheol.* **34**, 275–277.

Baker, C.T.H. 2000 A perspective on the numerical treatment of Volterra equations. *J. Comput. Appl. Math.* **2000**, 217–249.

Barnes, H.A. 1995 A review of the slip (wall depletion) of polymer solutions, emulsions and particle suspensions in viscometers: its cause, character, and cure. *J. Non-Newtonian Fluid Mech.* **56**, 221–251.

Barnes, H.A. 1997 Thixotropy – a review. *J. Non-Newtonian Fluid Mech.* **70**, 1–33.

Barnes, H.A. 1999 The yield stress – a review or '$\pi\alpha\nu\tau\alpha\ \rho\epsilon\iota$' – everything flows? *J. Non-Newtonian Fluid Mech.* **81**, 213–217.

Barnes, H.A. 2000 *A Handbook of Elementary Rheology*. Aberystwyth: University of Wales.

Barnes, H.A. & Carnali, J.O. 1990 The vane-in-cup as a novel rheometer geometry for shear thinning and thixotropic materials. *J. Rheol.* **34**, 851–866.

Barnes, H.A. & Nguyen, Q.D. 2001 Rotating vane rheometry – A review. *J. Non-Newtonian Fluid Mech.* **98**, 1–14.

Baudez, J.C., Rodts, S., Chateau, X. & Coussot, P. 2004 New technique for reconstructing instantaneous profiles from viscometric tests: application to pasty materials. *J. Rheol.* **48**, 69–82.

Bertero, M., De Mol, C. & Pike, E.R 1985 Linear inverse problems with discrete data. I: General formulation and singular system analysis. *Inverse Prob.* **1**, 301–330.

Bertero, M., De Mol, C. & Pike, E.R 1988 Linear inverse problems with discrete data. II: Stability and regularization. *Inverse Prob.* **4**, 573–594.

Bird, R.B., Armstrong, R.C. & Hassager, O. 1987 *Dynamics of Polymeric Liquids*. New York: John Wiley & Sons.

Bird, R.B., Dai, G.C. & Yarusso, B.J. 1983 The rheology and flow of viscoplastic materials. *Rev. Chem. Eng.* **1**, 1–70.

Bluman, G.W. 1990 Invariant solutions for ordinary differential equations. *SIAM J. Appl. Math.* **50**, 1706–1715.

Boehler, J.P., ed. 1987 *Applications of tensor functions in solid mechanics*, *International Centre for Mechanical Sciences, Courses and Lectures*, vol. 292. New York: Springer Verlag.

Boger, D.V. & Walters, K. 1993 *Rheological Phenomena in Focus*. Amsterdam: Elsevier.

Borgia, A. & Spera, F.J. 1990 Error analysis for reducing noisy wide-gap concentric cylinder rheometric data for nonlinear fluids: theory and applications. *J. Rheol.* **34**, 117–135.

Brady, J.F. & Morris, J.F. 1997 Microstructure of strongly sheared suspensions and its impact on rheology and diffusion. *J. Fluid Mech.* **348**, 103–139.

Cai, T.T. 1999 Adaptative wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Stat.* **27**, 898–924.

Cai, T.T. 2002 On adaptive wavelet estimation of a derivative and other related linear inverse problems. *J. Stat. Planning Infer.* **108**, 329–349.

Calvetti, D., Morigi, S., Reichel, L. & Sgallari, F. 2000 Tikhonov regularization and the L-curve for large discrete ill-posed problems. *J. Comput. Appl. Math.* **123**, 423–446.

Champier, S. & Grammont, L. 2002 A wavelet-vaguelet method for unfolding sphere size distributions. *Inverse Prob.* **18**, 79–94.

Coleman, B.D., Markowitz, H. & Noll, W. 1966 *Viscometric flows of non-Newtonian fluids*, *Springer Tracts in natural philosophy*, vol. 5. Berlin: Springer-Verlag.

Courant, R. & Friedrich, K.O. 1948 *Supersonic Flow and Shock Waves*. New York: Intersciences Publishers.

Darby, R. 1985 Couette viscometer data reduction for materials with a yield stress. *J. Rheol.* **29**, 369–378.

Dicken, V. & Maass, P. 1996 Wavelet-Galerkin methods for ill-posed problems. *J. Inv. Ill-Posed Problems* **4**, 203–221.

Donoho, D.L. 1995 Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comp. Harmonic Anal.* **2**, 101–126.

Donoho, D.L. & Johnstone, I.M. 1995 Wavelet shrinkage: asymptotia? *J. Royal. Stat. Soc. B* **57**, 301–369.

Donoho, D.L. & Johnstone, I.M. 1998 Minimax estimation via wavelet shrinkage. *Ann. Stat.* **26**, 879–921.

Doraiswamy, D., Mujumdar, A.N., Taso, I., Beris, A.N., Danforth, S.C. & Metzner, A.B. 1991 The Cox-Merz rule extended: a rheological model for concentrated suspensions and other materials with a yield stress. *J. Rheol.* **35**, 647–685.

Dressler, L. 1983 *Similarity Solutions of Nonlinear Partial Differential Equations*. Boston: Pitman.

Einstein, A. 1911 Berichtigung zu meiner Arbeit "Eine neue Bestimmung der Moleküldimensionen". *Annalen der Physik* **34**, 591–592.

Evans, I.D. 1992 On the nature of the yield stress. *J. Rheol.* **36**, 131–1318.

Frankel, N.A. & Acrivos, A. 1967 On the viscosity of a concentrated suspension of solid spheres. *Chem. Eng. Sci.* **22**, 847–853.

Friedrich, C., Honerkamp, J. & Weese, J. 1996 New ill-posed problems in rheology. *Rheol. Acta* **35**, 186–193.

Gamboa, F. & Gassiat, E. 1997 Bayesian methods and maximum entropy for ill-posed inverse problems. *Ann. Stat.* **25**, 328–350.

Geissle, W. & Hochstein, B. 2003 Validity of the Cox-Merz rule for concentrated suspensions. *J. Rheol.* **47**, 897–910.

Germain, P. 1973*a* La méthode des puissances virtuelles en mécanique des milieux continus. *Journal de Mécanique* **12**, 235–274.

Germain, P. 1973*b* The method of virtual power in continuum mechanics, Part 2: Microstructure. *SIAM J. Appl. Math.* **25** (3), 556–575.

Gilks, W.R., Richardson, S. & Spiegelhalter, D.J., ed. 1997 *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Golberg, M. 1979 A method of adjoints for solving some ill-posed equations of the first kind. *Appl. Math. Comput.* **5**, 123–130.

Golberg, M. 1995 A note on the sparse representation of discrete integral operators. *Appl. Math. Comput.* **71**, 97–118.

Harnett, J.P. & Hu, R.Y.Z. 1989 The yield stress–An engineering reality. *J. Rheol.* **33**, 671–679.

Hart, J.D. 1999 *Nonparametric Smoothing and Lack-of-Fit Tests*. New York: Springer.

Huppert, H.E. 1982 Flow and instability of a viscous current down a slope. *Nature* **300**, 427–429.

Hutter, K. & Jöhnk, K. 2004 *Continuum Methods of Physical Modeling*. Berlin: Springer.

Jones, C.W. 1953 On Reducible Non-Linear Differential Equations Occurring in Mechanics,. *Proc. R. Soc. London ser. A* **217**, 327–343.

de Kee, D. & Chan Man Fong, C.F. 1993 A true yield stress? *J. Rheol.* **37**, 775–776.

Kirsch, A. 1996 *An Introduction to the Mathematical Theory of Inverse Problems*. New York: Springer.

Kovac, A. 1998 Wavelet Thresholding for Unequally Spaced Data. Phd thesis, University of Bristol.

Krieger, I.M. 1968 Shear rate in the Couette viscometer. *Trans. Soc. Rheol.* **12**, 5–11.

Krieger, I.M. & Elrod, H. 1953 Direct determination of the flow curves of non-Newtonian fluids. II Shearing rate in the concentric cylinder viscometer. *J. Appl. Phys.* **24**, 134–136.

Krieger, I.M. & Maron, S.H. 1952 Direct determination of the flow curves of non-Newtonian fluids. *J. Appl. Phys.* **23**, 147–149.

Lacey, A.A., Ockendon, J. R. & Tayler, A. B. 1982 "Waiting-Time" Solutions of a Nonlinear Diffusion Equation. *SIAM J. Appl. Math.* **42**, 1252–1264.

Landau, D. & Binder, K. 2000 *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge: Cambridge University Press.

Leong, Y.K. & Yeow, Y.L. 2003 Obtaining the shear stress shear rate relationship and yield of liquid foods from Couette viscometry data. *Rheol. Acta* **42**, 365–371.

LeVeque, R.J. 2002 *Finite Volume Methods for Hyperbolic Problems*. Cambridge: Cambridge University Press.

Louis, A.K. 1996 Approximate inverse for linear and some nonlinear problems. *Inverse Prob.* **12**, 175–190.

Louis, A.K. 1999 A unified approach to regularization methods for linear and ill-posed problems. *Inverse Prob.* **15**, 489–498.

Louis, A.K. & Maass, P. 1990 A mollifier method for linear operator equations of the first kind. *Inverse Prob.* **6**, 427–440.

Louis, A.K., Maass, P. & Rieder, A. 1997 *Wavelets, Theory and Applications*. Chichester: John Wiley & Sons.

MacKay, D.J.C. 2003 *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Macsporran, W.C. 1989 Direct numerical evaluation of shear rates in concentric cylinder viscometry using least-squares cubic splines. *J. Rheol.* **33**, 745–755.

Magnin, A. & Piau, J.M. 1987 Shear rheometry of fluids with a yield stress. *J. Non-Newtonian Fluid Mech.* **23**, 91–106.

Magnin, A. & Piau, J.M. 1990 Cone-and-plate rheometry of yield stress fluids. Study of an aqueous gel. *J. Non-Newtonian Fluid Mech.* **36**, 85–108.

Mallat, S. 1998 *A Wavelet Tour of Signal Processing*, 2nd edn. San Diego: Academic Press.

McKarthy, K. & Seymour, J. 1994 Gravity current analysis of the Bostwick consistometer for power-law foods. *Journal of Texture Studies* **25**, 207–220.

Mendes, P.R.S. & Dutra, E.S.S. 2004 Viscosity function for yield-stress liquids. *Applied Rheology* **14**, 296–302.

Middleton, G.V. & Wilcock, P.R. 1994 *Mechanics in the Earth and Environmental Sciences*. Cambridge: Cambridge University Press.

Minier, J.-P. & Peirano, E 2001 The pdf approach to turbulent polydispersed two-phase flows. *Physics Reports* **352** (1-3), 1–214.

Mooney, M. 1931 Explicit formulas for slip and fluidity. *J. Rheol.* **2**, 210–222.

Morrison, F.A. 2001 *Understanding Rheology*. New York: Oxford University Press.

Mosegaard, K. & Sambridge, M. 2002 Monte Carlo analysis of inverse problems. *Inverse Prob.* **18**, R29–R54.

Nguyen, Q.D. & Boger, D.V. 1992 Measuring the flow properties of yield stress fluids. *Annu. Rev. Fluid Mech.* **24**, 47–88.

Oldroyd, J.G. 1950 On the formulation of rheological equations of state. *Proc. R. Soc. London ser. A* **200**, 523–541.

O'Sullivan, F. 1986 A statistical perpective on ill-posed inverse problems. *Stat. Sci.* **1**, 502–527.

Pashias, N. & Boger, D.V. 1996 A fifty cent rheometer for yield stress measurement. *J. Rheol.* **40**, 1179–1189.

Piau, J.-M. & Debiane, K. 2005 Consistometers rheometry of power-law viscous fluids. *J. Non-Newtonian Fluid Mech.* **127**, 213–224.

Pope, S.B. 2000 *Turbulent Flows*. Cambridge: Cambridge University Press.

Rieder, A. 1997 A wavelet multilevel method for ill-posed problems stabilized by Tikhonov regularization. *Numer. Math* **75**, 501–522.

Robert, C. 2001 *The Bayesian Choice*. New York: Springer.

Roberts, G.P. & Barnes, H.A. 2001 New measurements of the flow-curves for Carbopol dispersions without slip artefacts. *Rheol. Acta* **40**, 499–503.

Schowalter, W.R. & Christensen, G. 1999 Toward a rationalization of the slump test for fresh concrete: comparisons of calculations and experiments. *J. Rheol.* **42**, 865–870.

Shaqfeh, E.S.G. 1996 Purely elastic instabilities in viscometric flows. *Annu. Rev. Fluid Mech.* **28**, 129–185.

Spaans, R.D. & Williams, M.C. 1995 At last, a true liquid-phase yield stress. *J. Rheol.* **39**, 241–246.

Tanner, M.A. 1996 *Tools for Statistical Inference, Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd edn. New York: Springer.

Tanner, R.I. 1988 *Engineering Rheology*. Oxford: Clarendon Press.

Tanner, R.I. & Williams, G. 1970 Iterative numerical methods for some integral equations arising in rheology. *Trans. Soc. Rheol.* **14**, 19–38.

Tenorio, L. 2001 Statistical regularization of inverse problems. *SIAM Rev.* **43**, 347–366.

Truesdell, C. 1966 *Six Lectures on Modern Natural Philosophy*. New York: Springer Verlag.

Truesdell, C. 1974 The meaning of viscometry in fluid dynamics. *Annu. Rev. Fluid Mech.* **6**, 111–147.

Vidakovic, B. 1999 *Statistical Modeling by Wavelets*. New York: John Wiley & Sons, Inc.

Wahba, G. 1990 *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

Werman, M. & Keren, D. 2001 A Bayesian method for fitting parametric and nonparametric models to noisy data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**, 528–534.

Wieghardt, K. 1975 Experiments in granular flow. *Annu. Rev. Fluid Mech.* **7**, 89–114.

Yang, T.M.T. & Krieger, I.M. 1978 Comparison of methods for calculating shear rates in coaxial viscometers. *J. Rheol.* **22**, 413–421.

Ycart, B. 2002 *Modèles et Algorithmes Markoviens*. Berlin: Springer.

Yeow, Y.L., Ko, W.C. & Tang, P.P.P. 2000 Solving the inverse problem of Couette viscometry by Tikhonov regularization. *J. Rheol.* **44**, 1335–1351.

Zarraga, I.E., Hill, D.A. & Leighton, D.T. 2000 The characterization of the total stress of concentrated suspensions of noncolloidal spheres in Newtonian fluids. *J. Rheol.* **44**, 185–221.

Zdunkowski, W. & Bott, A. 2003 *Dynamics of the Atmosphere*. Cambridge: Cambridge University Press.

Zheng, Q.-S. 1994 Theory of representations for tensor functions- A unified invariant approach to constitutive equation. *Appl. Mech. Rev.* **47**, 545–586.

Index