# Introduction à la modélisation

G. Barles

23 novembre 2015

## Table des matières

# 1 Introduction

Le but de ce cours est de décrire quelques aspects de la modélisation de phénomènes essentiellement déterministes mais avant cela, on peut légitimemement se poser la question :

*Qu'est-ce que la modélisation ? Quel est son but ?*

Donner une définition précise de la modélisation n'est pas chose aisée car elle peut apparaître sous différentes formes suivant les domaines ou les phénomènes à modéliser. Nous en verrons plus loin quelques exemples assez différents. Une tentative de définition (à affiner !) pourrait être : l'objectif de la modélisation, c'est d'écrire un certain nombre d'équations mathématiques dont les solutions décrivent "assez bien" le phénomène auquel on s'intéresse, dans le but de prédire l'évolution d'un système à partir d'une situation donnée ou de mieux comprendre les raisons de cette évolution.

En général, sur une problématique donnée, la première modélisation vise à comprendre les causes du phénomène : le modèle doit donc être suffisamment simple pour pouvoir être analysé complètement et savoir quel(s) terme(s) dans l'équation provoque tel ou tel effet. Ensuite, une fois convaincu de la pertinence de cette première modélisation, on peut la compliquer pour prendre en compte des aspects de plus en plus sophistiqués.

Ceci crée des différences entre les disciplines : dans le cas des sciences physiques ou des sciences pour l'ingénieur, la modélisation s'appuie sur des siècles de pratique et d'expérience ainsi que sur l'existence d'un nombre important de lois physiques bien établies ; on s'intéresse aussi à des phénomènes clairement déterministes, les mêmes causes induisants les mêmes effets, inéluctablement.

À cause de ces caractéristiques, dans de nombreuses situations voire la plupart, la modélisation est vraiment utilisée dans le but de prédire le comportement de tel ou tel système (écoulement autour d'une aile d'avion, transition de phase, résistance des matériaux, trajectoire de fusée...). Et il est clair aussi que la modélisation se révèle être un outil précieux sinon indispensable pour les industriels.

Dans le cas d'autres sciences (biologie, économie,...), l'intervention des êtres vivants perturbe un peu le déterminisme, même si on espère conserver via une compréhension plus fine des phénomènes, un certain aspect prévisible, au moins statistiquement. En tout cas, l'aspect historique est complètement différent car le souci de modélisation (en tout cas à un certain niveau de complexité) est assez récent dans le temps. Dans ces disci-

plines, on doit mettre en évidence des “lois de comportement” qui s'appuie souvent sur des méthodes de modélisation “traditionnelles” mais on observe un peu plus l'intervention d'aspects aléatoires, déjà présent en mécanique statistique. L'exemple de la mécanique statistique montre bien comment l'aléatoire peut conduire à des phénomènes déterministes si on a un effet de moyennisation.

Ces sciences où l'intervention de la modélisation est plus récentes (ou du moins avec une approche scientifique plus récente = moins d'un siècle) nous ramènent loin en arrière, à ce qu'ont certainement été les premiers pas en physique mais aussi dans des contextes assez différents : par exemple, dans le cas des sciences humaines, on peut parfois se demander si une quelconque modélisation est possible tant l'intervention irraisonnée de la psychologie humaine peut avoir des conséquences tout à fait imprévisibles. Pensons, par exemple, aux situations de panique à la bourse qui sont souvent disproportionnées par rapport à la situation économique réelle. Parmi ces sciences plus récentes où la modélisation a quelquefois été une conséquence de l'apparition et des progrès de l'informatique, on peut évoquer les sciences économiques, parmi lesquelles la finance, les sciences d'aides à la décision (théorie des jeux, apprentissage...), l'informatique et le traitement d'images.

Ce dernier exemple très à la mode recouvre beaucoup de problématiques différentes mais avec un point commun important : il ne s'agit pas ici de prédire mais de comprendre comment structurer et/ou traiter une image. Une image ou son traitement doit ressembler à quel phénomène physique ? quels sont ses constituants essentiels ? A cause de cela, l'analyse d'images fait appel à des modèles physiques et mathématiques très sophistiqués aussi bien pour le débruitage (amélioration de la qualité d'une image) que pour la simplification d'image (détection de contour...).

*Comment modéliser ? Quelles sont les étapes principales ?*

Malheureusement (ou heureusement ?), la diversité des situations décrites ci-dessus empêche l'existence de règles uniformes. Mais il existe un certain nombre de passages obligés.

La première difficulté, c'est d'identifier clairement les **variables** et les **paramètres** importants du système considéré ; mathématiquement cela signifie : quelles sont les inconnues pertinentes que l'on va utiliser dans nos équations et de quelles variables dépendent-elles ? Quels sont les paramètres additionnels ? En d'autres termes, quels sont les facteurs explicatifs du phénomène considéré et quelles sont les grandeurs que l'on estime essentielles ?

La deuxième étape consiste à trouver des relations entre ces inconnues,

ces variables et ces paramètres (**lois de comportement**) ainsi que des **équations régissant l'évolution des inconnues**.

Cette deuxième étape conduit très souvent à un nombre d'équations trop important ou à des équations trop complexes. Il est donc nécessaire de simplifier ces équations en analysant l'importance des différents termes, ce qui amène à supprimer des inconnues, des variables, des paramètres et des équations. La question est : **peut-on négliger tel ou tel phénomène ? sait-on a priori que telle ou telle quantité est "petite" et donc on peut la supposer nulle ?** Bien entendu, on peut se tromper et revenir en arrière. C'est en grande partie cela le travail de modélisation car plus un modèle est simple, plus il est utilisable et par là même utile.

Ensuite, quand c'est possible, on procède à une étude mathématique des équations restantes. Evidemment, une situation idéale est celle où ces équations ont des solutions totalement explicites, mais c'est rare : dans ce cas, on a un accès direct aux solutions mais aussi (et surtout) à leurs dépendances en les différentes variables et les divers paramètres. Ceci permet des études simples de sensibilités et donne une compréhension plus aisée des phénomènes.

Malheureusement, dans beaucoup de cas, on ne sait pas dire grand chose mais cette étude doit au moins analyser le type des équations considérées et d'en déduire les méthodes numériques adaptées pour "calculer numériquement" la (ou les) solution(s) avec efficacité. Car on ne sait bien calculer numériquement que les solutions d'équations pour lesquelles on a un peu de théorie (au moins qualitative) et encore...

En général, on procède soit avant l'étude mathématique, soit au moment de l'implémentation numérique à **l'adimensionalisation** des équations. Les ordinateurs ne gérant vraiment bien que les quantités entre 1 et 10, on essaie de faire en sorte que les quantités considérées soient dans cet intervalle et pour cela, on les normalise par une quantité de référence. Exemple : si on doit calculer des longueurs $l$ de l'ordre de $10^4$ mètres, on introduira $\tilde{l} = 10^{-4} l$ avec des chances pour que $\tilde{l}$ soit de l'ordre de 1. Si on s'intéresse à des fusées ou des escargots, l'adimensionalisation de la vitesse ne se fera pas sur les mêmes bases ! De manière générale, si $x$ est une variable ou une fonction à calculer, on s'intéressera plutôt à :

$$\tilde{x} = \frac{x}{\bar{x}} \, ,$$

où $\bar{x}$ est soit une valeur moyenne de $x$, soit une valeur caractéristique de $x$.

Ensuite, il faut être prudent quant à l'interprétation des simulations

numériques : les méthodes numériques peuvent générer des arte fact de calcul n'ayant rien à voir avec la réalité... cela a conduit pendant quelques années des patients sur le billard parce que l'image "montrait" une région suspecte (avec possibilité de tumeur cancéreuse) . Une remarque essentielle : en général, on teste les méthodes numériques sur des exemples simples puis on les utilise dans des cas plus complexes. On ne fait jamais confiance à une méthode numérique les yeux fermés.

*Validation et limites de la modélisation*

Une fois la simulation réalisée, on compare (en général) ses résultats avec l'observation ; cette deuxième étape permet de juger la qualité de la modélisation et elle conduit, éventuellement, à la modifier : le processus de modélisation est, le plus souvent, un processus itératif (100 fois sur le métier remettez votre ouvrage...).

La question des limites des modèles est trop souvent négligée par les utilisateurs et même par les scientifiques eux-mêmes : quelles ont été les hypothèses utilisées pour élaborer le modèle ? Sont-elles satisfaites dans la situation considérée ?

Il faut aussi faire attention au fait qu'un bon ajustement à l'expérience ne prouve rien, par exemple quand le nombre de mesures est insuffisant. De plus, lors de la comparaison modèle-expérience, il faut prendre en compte les erreurs de mesure. Dans le cadre de phénomènes aléatoires, un grand nombre d'expérience est nécessaire pour échapper aux cas exceptionnels. Dans tous les cas, il faut être conscient que seule la répétition d'un nombre important d'expériences aux conclusions concordantes valide de façon pragmatique (et temporaire) un modèle.

Dans certains domaines d'observation, il n' y a pas d'expérience directe possible, il faut alors observer non le phénomène lui-même mais ses conséquences . C'est le cas notamment dans le domaine de la relativité et de l'astronomie (effet Doppler pour constater l'expansion de l'univers).

D'autres difficultés peuvent apparaître, à commencer par la sensibilité aux conditions initiales ; par exemple lorsque l'on veut étudier la trajectoire d'une feuille de papier. Le fameux effet "aile de papillon" ! Cela se produit aussi quand le modèle n'assure pas une solution mathématique unique : le fait de simuler le processus avec une donnée précise ou même avec plusieurs données initiales ne renseigne absolument pas sur ce qui va se passer avec une donnée initiale voisine !

Pour d'autres raisons, certains phénomènes ne sont pas non plus prédictibles

comme en sciences économiques, où les modèles complexes sont (ou devraient) être utilisés pour mieux comprendre les phénomènes, pour donner une tendance et pas de manière prédictive ou alors en étant bien conscient des limites d'un tel modèle.

Enfin, des problèmes purement techniques peuvent compliquer les simulations numériques : puissance de calcul trop faible (météo) ou systèmes linéaires mal déterminés (ils n'ont de solutions que sous certaines conditions).

Dans ce cours, nous allons examiner trois types de situations en fonctions du type de problématiques mathématiques à laquelle on aboutit après la modélisation : une première partie où les modèles obtenus sont des équations différentielles ordinaires, une seconde où l'on obtient des équations aux dérivées partielles et une dernière où le thème central sera l'optimisation.

# 2 Modélisation et équations différentielles ordinaires

## 2.1 Un peu de physique du point matériel

Dans beaucoup de situations physiques, l'objet d'étude peut être assimilé à un point matériel : planète dans le système solaire, bille sur un plan incliné, l'hypothèse "point matériel" provient soit de la taille "petite" (relativement), soit du fait que seul le comportement du centre de gravité du corps considéré joue un rôle.

En Mécanique du point matériel, une loi fondamentale est le principe fondamental de la dynamique que l'on peut résumer par la formule :

$$F = m\gamma \ ,$$

où $F$ désigne la "résultante" (la somme) des forces extérieures, $m$ la masse du point matériel et $\gamma$ son accélération. Si on note $x(t)$ la position du point matériel à l'instant $t$, alors $\gamma = \ddot{x}(t)$ où " $\dot{}$ " désigne la dérivée par rapport à $t$ et donc ici " $\ddot{}$ " est une dérivée seconde.

Les forces les plus communes sont la gravitation, les forces électro-statiques et on dit que $F$ dérive d'un potentiel si :

$$F = -\nabla U \ ,$$

où $U$ est le potentiel. Ceci nous amène à parler d'énergie car si on est dans le cas d'une force dérivant d'un potentiel, le principe fondamental de la dynamique nous dit que :

$$m\ddot{x}(t) = -\nabla U(x(t)) \ ,$$

et on voit facilement (en dérivant) que :

$$E(t) = \frac{1}{2}m(\dot{x}(t))^2 + U(x(t)) = \text{constante} .$$

La quantité constante ("conservé") $E(t)$ est l'énergie du système, le premier terme désignant l'énergie cinétique et le second l'énergie potentielle.

## 2.2 L'exemple du pendule

Pour mettre en application la physique du point matériel, on s'intéresse au pendule simple qui est une masse ponctuelle fixée à l'extrémité d'un fil sans masse, inextensible et sans raideur, lui même ayant son autre extrémité fixée en un point $O$. Il oscille sous l'effet de la pesanteur. Si sa masse est normalisée à $m = 1$, l'équation s'écrit :

$$\ddot{x}(t) = \overrightarrow{g} + \overrightarrow{R} ,$$

où $\overrightarrow{g}$ est la force de pesanteur, $\overrightarrow{R}$ est la tension du fil (une force qui exprime que la masse ponctuelle est attachée au fil et donc ne tombe pas) et $x(t)$ est la position de la masse ponctuelle à l'instant $t$.

On suppose que le modèle est écrit dans $\mathbb{R}^2$ et il s'agit d'écrire les équations de manière plus utilisable. Pour cela, on prend $O$ comme origine et on utilise des coordonnées polaires : si $r$ est la longueur du fil alors :

$$x(t) = r\exp(i\theta(t)) .$$

On peut faire les calculs soit en repère mobile soit via les nombres complexes : dans ce dernier cas, on a :

$$\dot{x}(t) = i\dot{\theta}(t)r\exp(i\theta(t)) \quad \text{et } \ddot{x}(t) = i\ddot{\theta}(t)r\exp(i\theta(t)) - (\dot{\theta}(t))^2 r\exp(i\theta(t)) .$$

Le premier terme dans le calcul de $\ddot{x}(t)$ est l'accellération tangentielle (dans la direction orthogonale au fil $i\exp(i\theta(t))$) et le second terme est la force centrifuge (dans la direction radiale $\exp(i\theta(t))$).

Comme il n'y a claiment pas de mouvement dans la direction radiale (le poids, la tension du fil et la force centrifuge s'équilibrant parfaitement), il reste la direction tangentielle où seule la résultante de la gravité intervient : si l'axe $\theta = 0$ est la verticale dirigée vers le centre de la terre, on a :

$$\overrightarrow{g} = g(\in \mathbb{R}) = g\cos(\theta(t))\exp(i\theta(t)) - g\sin(\theta(t))i\exp(i\theta(t)) \text{ (1)} ,$$

et donc :

$$\ddot{\theta}(t)r = -g\sin(\theta(t)) .$$

(1). Il est à noter ici que l'on écrit tout dans le repère mobile formé des vecteurs $\exp(i\theta(t))$ et $i\exp(i\theta(t))$.

En normalisant les constantes à 1 ($r = g = 1$), on a donc une équation différentielle ordinaire simple :

$$\ddot{\theta}(t) = -\sin(\theta(t)) \ ,$$

qui nécessite la connaissance de (par exemple) $\theta(0)$, $\dot{\theta}(0)$ pour être résolue.

Quelques remarques :

1. Il s'agit d'une équation non linéaire qui n'est pas évidente à résoudre...
2. En multipliant par $\dot{\theta}(t)$ et en intégrant, on voit que :

$$\frac{1}{2}(\dot{\theta}(t))^2 - \cos(\theta(t)) = \text{constante} \ .$$

   On a affaire à un système "conservatif" où l'énergie est préservée, le premier terme étant, comme ci-dessus, un terme d'énergie cinétique et le second un terme d'énergie potentielle.
3. Nous avons fait beaucoup d'hypothèses simplificatrices pour arriver à ce modèle élémentaire : point matériel, absence de frottements (en pratique, le pendule s'arrête... ici, non), la terre est plate,...etc.

## 2.3 Étude mathématique : problèmes et solutions ?

On va dans cette section regarder plusieurs propriétés des solutions de cette équation différentielle.

**Théorème 1. (Existence, unicité, stabilité (I))** *Pour toute donnée initiale $\theta(0) = a$, $\dot{\theta}(0) = b$, l'équation du pendule a une unique solution définie pour tous temps. De plus cette solution dépend continûment de $a$ et $b$.*

*Démonstration.* Un bon cours d'edo et aucune difficulté... □

Une attention est toujours portée sur les points d'équilibre, c'est-à-dire ceux pour lesquels la solution ne bouge pas. Pour le pendule, cela semble assez évident mais en résolvant $\dot{\theta}(t) = \ddot{\theta}(t) = 0$ pour tout $t$, on trouve deux équilibre : $(0, 0)$ et $(\pi, 0)$ (modulo $2\pi$).

L'étude de la stabilité de ces points d'équilibre est une première occasion de faire intervenir une méthode très utilisée : la **linéarisation**. Si on connait une trajectoire $\theta(\cdot)$ (par exemple explicitement), on cherche des solutions proches en considérant $\theta(\cdot) + h(\cdot)$ où $h(\cdot)$ est supposée "petite" (par exemple pour la norme du sup ou la norme $C^1$). Ceci conduit à écrire :

$$\ddot{\theta}(t) + \ddot{h}(t) = -\sin(\theta(t) + h(t)) = -\sin(\theta(t)) - \cos(\theta(t))h(t) + \cdots \ .$$

Si on utilise l'équation satisfaite par $\theta$ et si on néglige les $\cdots$ en ne conservant que le terme "linéaire", on aboutit à :

$$\ddot{h}(t) = -\cos(\theta(t))h(t) \ .$$

Si on utilise la solution d'équilibre $\theta(t) \equiv 0$, on est amené à considérer l'edo $\ddot{h}(t) = -h(t)$ dont les solutions sont des sin et cos et on conclut à la stabilité. Par contre, pour la solution d'équilibre $\theta(t) \equiv \pi$, on est amené à considérer l'edo $\ddot{h}(t) = h(t)$ dont les solutions sont $\exp(t)$ et $\exp(-t)$, la première étant non bornée pour les $t > 0$.

Un modèle plus général prend en compte un terme de frottement qui s'oppose au mouvement, par exemple en $-\alpha\dot{x}(t)$. On vérifie facilement que la nouvelle équation est de la forme :

$$\ddot{\theta}(t) = -\alpha\dot{\theta}(t) - \sin(\theta(t)) \ ,$$

puis on prouve de la même manière l'existence, l'unicité et la stabilité des solutions : on a même des résultats de stabilité meilleurs, ce qui ne doit pas surprendre du point de vue physique ?

Il reste une question à aborder : si on ne connait pas les solutions explicitement, on fait comment ? On construit des **schémas numériques.**

Pour ce faire, on écrit l'équation sous la forme :

$$\dot{y}(t) = f(t, y(t)) \ ,$$

où, ici, $y(t) = (\theta(t), \dot{\theta}(t))$ et $f(t, (y_1, y_2)) := (y_2, -\sin(y_1))$.

Pour simplifier la présentation, on utilisera l'hypothèse simplifiée suivante ($f$ est "globalement lipschitzienne" en $y$) :

**(GL)** La fonction $f : [0, T] \times \mathbb{R}^n \to \mathbb{R}^n$ est continue et il existe une constante $L$ telle que :

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2| \ ,$$

pour tous $t \in [0, T]$ et $y_1, y_2 \in \mathbb{R}^n$.

## 2.4 Rappels sur les schémas numériques pour les EDO

On va d'abord présenter la méthode d'Euler.

Pour résoudre numériquement l'EDO, on va se donner une *grille*, c'est-à-dire des points $t_0 = 0 < t_1 < t_2 < t_3 < \cdots < t_N = T$, et on va essayer de calculer une "bonne" approximation des valeurs de la solution en tous ces points, c'est-à-dire des valeurs $y_i \simeq y(t_i)$.

Numériquement le sens de "bonne approximation" n'est pas absolu car, d'une part, l'ordre de grandeur joue un rôle (une approximation à 100 000 ans près peut être excellente si on est géologue...) et, d'autre part, le temps mis pour calculer la solution peut être un facteur important (quel est l'intérêt d'avoir un résultat très précis s'il faut 10 ans pour l'obtenir ?). Il y a toujours,

dans les méthodes numériques un “rapport qualité - prix ” : précision vs temps de calcul ou complexité.

Essentiellement il y a deux approches pour calculer les $y_i$ qui, dans le cas de la méthode d'Euler, vont aboutir au même résultat : soit on approche directement l'EDO par “différences finies” en utilisant une approximation de la dérivée $y'(t)$, soit on intègre l'EDO, se rapprochant ainsi de la preuve d'existence.

L'approche par *“différences finies”* consiste à approcher $y'(t_i)$ par différences finies ; par exemple :

$$y'(t_i) \simeq \frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i} \ .$$

L'EDO se réécrit alors sous la forme :

$$\frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i} \simeq f(t_i, y(t_i)) \ ,$$

et donc :

$$y(t_{i+1}) \simeq y(t_i) + (t_{i+1} - t_i) f(t_i, y(t_i)) \ .$$

Ceci suggère que l'on peut calculer les $y_k$ via la relation de récurrence :

$$y_{i+1} = y_i + (t_{i+1} - t_i) f(t_i, y_i) \ ,$$

le terme $y_0$ étant connu (donnée initiale). C'est la *méthode d'Euler*.

La deuxième approche, qui va nous conduire au même résultat mais avec une philosophie très différente, consiste à intégrer l'équation de $t_i$ à $t_{i+1}$ :

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(s, y(s)) ds \ .$$

Si on applique la méthode des rectangles à l'intégrale de la manière suivante :

$$\int_{t_i}^{t_{i+1}} f(s, y(s)) ds \simeq (t_{i+1} - t_i) f(t_i, y(t_i)) \ ,$$

on retrouve les calculs ci-dessus et la méthode d'Euler.

**Remarque :** En pratique, il peut être intéressant d'utiliser une subdivision adaptée avec plus de points aux endroits où la fonction $y$ varie beaucoup et moins de points aux endroits où les variations sont faibles. Mais choisir une telle subdivision (ou faire en sorte que l'ordinateur choisisse automatiquement cette subdivision = schémas adaptatifs) n'est pas toujours simple.

### 2.4.1 Étude de l'erreur (I)

Désormais nous nous plaçons dans le cadre d'une grille uniforme :

$$t_{i+1} - t_i = h = \frac{T}{N} \ .$$

La méthode d'Euler s'écrit alors :

$$y_{i+1} = y_i + hf(t_i, y_i) \ .$$

On notera :

$$e_i = y(t_i) - y_i \ ,$$

l'erreur commise au point $t_i$ et :

$$\varepsilon_i = y(t_{i+1}) - y(t_i) - hf(t_i, y(t_i)) \ ,$$

l'erreur de "consistance" ; c'est l'erreur systématique commise sur $y_{i+1}$ : même si on a calculé exactement la valeur à l'instant $t_i$ ($y_i \simeq y(t_i)$), on a une erreur sur $y(t_{i+1})$ qui est $\varepsilon_i$.

Pour évaluer l'erreur, on procède comme suit :

$$\begin{aligned} e_{i+1} &= y(t_{i+1}) - y_{i+1} \\ &= [y(t_{i+1}) - y(t_i) - hf(t_i, y(t_i))] + [y(t_i) - y_i] + [y_i + hf(t_i, y_i)] + \\ &\quad h[f(t_i, y(t_i)) - hf(t_i, y_i)] - y_{i+1} \\ &= \varepsilon_{i+1} + e_i + h[f(t_i, y(t_i)) - hf(t_i, y_i)] \ . \end{aligned}$$

D'où :

$$\begin{aligned} |e_{i+1}| &\leq |\varepsilon_{i+1}| + |e_i| + h|f(t_i, y(t_i)) - hf(t_i, y_i)| \\ &\leq |\varepsilon_{i+1}| + (1 + Lh)|e_i| \ , \end{aligned}$$

en utilisant le caractère lipschitz de $f$.

Il reste à estimer $|\varepsilon_{i+1}|$. En utilisant l'équation, on a :

$$\begin{aligned} |\varepsilon_{i+1}| &= |y(t_{i+1}) - y(t_i) - hf(t_i, y(t_i))| \\ &= |\int_{t_i}^{t_{i+1}} y'(s)ds - \int_{t_i}^{t_{i+1}} y'(t_i)ds| \\ &= |\int_{t_i}^{t_{i+1}} [y'(s) - y'(t_i)]ds| \\ &\leq h \sup_{s \in [t_i, t_{i+1}]} |y'(s) - y'(t_i)| \\ &\leq h.\omega(h, y') \ , \end{aligned}$$

où $\omega(\cdot, y')$ est le module de continuité de la fonction (continue) $y'$ sur $[0, T]$.

Finalement on a l'estimation suivante de l'erreur $e_{i+1}$ :

$$|e_{i+1}| \leq (1 + Lh)|e_i| + h.\omega(h, y') \ .$$

Pour conclure on utilise le :

**Lemme 1. (Lemme de Gronwall discret)**
*Si $(\theta_i)_i$ est une suite de réels positifs qui satisfait :*

$$\theta_{i+1} \leq (1+A)\theta_i + B \ ,$$

*où $A, B$ sont des constantes strictement positives alors :*

$$\theta_i \leq \exp(iA)\theta_0 + \frac{\exp(iA)-1}{A} B \ .$$

On utilise d'abord le lemme avec $A = Lh$ et $B = h.\omega(h, y')$ :

$$|e_i| \leq \exp(Lih)|e_0| + \frac{\exp(Lih)-1}{Lh} h.\omega(h, y') \ .$$

Mais $ih = t_i$ et (a priori) $e_0 = 0$ [ou du moins, $e_0$ est très petit] donc :

$$|e_i| \leq \frac{\exp(Lt_i)-1}{L}\omega(h, y') \leq \frac{\exp(LT)-1}{L}\omega(h, y') \ .$$

On vient donc de prouver le :

**Théorème 2.** *Sous l'hypothèse (GL) :*

$$\max_{0\leq i\leq N} |e_i| \leq \frac{\exp(LT)-1}{L}\omega(h, y') \ .$$

*En particulier,* $\max_{0\leq i\leq N} |e_i| \to 0$ *quand* $N \to +\infty$.

On donne maintenant la :
**Preuve du Lemme de Gronwall discret :** On note $(u_i)_i$ la suite définie par :

$$u_i = \frac{\theta_i}{(1+A)^i} \ .$$

En divisant la propriété satisfait par la suite $(\theta_i)_i$ par $(1+A)^{i+1}$, on voit que :

$$u_{i+1} \leq u_i + \frac{B}{(1+A)^{i+1}} \ .$$

Une récurrence immédiate montre que :

$$u_i \leq u_0 + \sum_{k=0}^{i-1} \frac{B}{(1+A)^{k+1}} = u_0 + \frac{B}{1+A}\frac{1-a^i}{1-a} \ ,$$

où $a = 1/(1+A)$. Comme :

$$\frac{1}{1-a} = \frac{1+A}{A} \ ,$$

il en résulte :

$$u_i \leq u_0 + B\frac{1-a^i}{A} \ ,$$

et donc :

$$\theta_i \leq (1+A)^i u_0 + B\frac{(1+A)^i - 1}{A} \ .$$

Il reste à remarquer que $1 + A \leq \exp(A)$, ce qui est clair puisque :

$$\exp(A) = 1 + A + \frac{A^2}{2!} + \cdots + \frac{A^n}{n!} + \cdots \ .$$

□

À titre d'exercice, on pourra démontrer la :

**Proposition 1.** *Si $y_h$ est la fonction affine par morceaux telle que $y_h(t_i) = y_i$ pour tout $i$, on a :*

$$||y_h - y||_\infty \to 0 \quad \textit{quand } h \to 0 \ .$$

### 2.4.2 Étude de l'erreur (II)

Le section précédente donne une estimation de convergence qui dépend du module de continuité de $y'$. Mais la fonction $y$ est inconnue donc ce résultat n'est pas satisfaisant car il n'est pas explicite. Nous allons maintenant donner une autre estimation qui ne dépend que des données, c'est-à-dire de $f$.

Pour cela, on étudie le module de continuité de $y'$ :

$$\begin{aligned} |y'(t) - y'(s)| &= |f(t, y(t)) - f(s, y(s))| \\ &= |f(t, y(t)) - f(s, y(t)) + f(s, y(t)) - f(s, y(s))| \\ &\leq |f(t, y(t)) - f(s, y(t))| + |f(s, y(t)) - f(s, y(s))| \\ &\leq |f(t, y(t)) - f(s, y(t))| + L|y(t) - y(s)| \end{aligned}$$

D'après les résultats sur les équations différentielles, $|y(t)| \leq D$ pour une certaine constante $D$ sur l'intervalle $[0, T]$ et le premier terme est estimé par le module de continuité de $f$ sur $[0, T] \times \overline{B}(0, D)$, noté $\omega_D(\cdot, f)$.

Quant au second, par le Théorème des Accroissements Finis :

$$|y(t) - y(s)| \leq M_f |t - s| \ ,$$

où :

$$M_f = \max_{[0,T]\times\overline{B}(0,D)} |f(t, y)| \ .$$

Finalement :

$$\omega(h, y') \leq \omega_D(h, f) + LM_f h \ ,$$

ce qui donne le résultat suivant qui était notre objectif :

**Théorème 3.** *Sous l'hypothèse (GL) :*

$$\max_{0\leq i\leq N} |e_i| \leq \frac{\exp(LT)-1}{L}(\omega_D(h,f)+LM_f h) \ .$$

On renvoit à l'appendice pour l'étude générale des méthodes à un pas qui s'écrivent

$$\left\{ \begin{array}{rcl} y_{i+1} & = & y_i + h\Phi(t_i, y_i, h) \\ y_0 & = & y_{0,h} \end{array} \right.$$

où $\Phi$ est une fonction continue sur $[0,T]\times\mathbb{R}^n\times[0,H]$, $H$ désignant un pas de discrétisation maximal.

### 2.4.3 Quelques élément sur les méthodes de Runge-Kutta

Ces méthodes sont les plus utilisées : elles sont rentrées "en standard" dans la plupart des logiciels. Comment marchent-elles ?

On repart de l'idée fondamentale qui consiste à écrire :

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(s, y(s))ds \ .$$

Nous avons vu dans la deuxième partie que pour calculer une intégrale, on utilise une formule de quadrature :

$$\int_0^1 \psi(s)ds \simeq \sum_{j=0}^{q} b_j \psi(c_j) \ ,$$

où $c_0 < c_1 < \cdots < c_q$, ce qui, en prenant $\psi(s) = f(t_i + sh, y(t_i + sh))$, nous donne :

$$y(t_{i+1}) \simeq y(t_i) + h\sum_{j=0}^{q} b_j f(t_{i,j}, y(t_{i,j})) \ ,$$

où $t_{i,j} = t_i + c_j h$. Ceci suggère une méthode que l'on peut écrire :

$$y_{i+1} = y_i + h\sum_{j=0}^{q} b_j k_{i,j} \ ,$$

où $k_{i,j}$ est une approximation de $f(t_{i,j}, y(t_{i,j}))$

Le problème, c'est qu'il faut encore calculer des approximations $y_{i,j}$ des $y(t_{i,j})$ pour avoir celle de $k_{i,j}$ et ceci est fait via :

$$y_{i,j} = y_i + h\sum_{k=0}^{q} a_{j,k} f(t_{i,k}, y(t_{i,k})) \ .$$

Cette procédure a l'air d'induire des équations non-linéaires couplées difficiles à résoudre et pour que ce ne soit pas le cas, on suppose que les $y_{i,j}$ ne

dépendent que des points déjà calculés, c'est-à-dire des $y_{i,k}$ pour $k < j$. On a donc :

$$y_{i,j} = y_i + h \sum_{k=0}^{j-1} a_{j,k} f(t_{i,k}, y(t_{i,k}))\ ,$$

et les $y_{i,j}$, ainsi que les $k_{i,j}$, sont calculés de proche en proche.

On résume souvent une méthode de Runge-Kutta grâce à un tableau de la forme :

$$\begin{array}{c|ccc} c_1 & a_{1,1} & \cdots & a_{1,q} \\ \vdots & \vdots & \cdots & \vdots \\ c_q & a_{q,1} & \cdots & a_{q,q} \\ \hline & b_1 & \cdots & b_q \end{array}$$

**Exemples :**

• $q = 1$ : C'est la méthode d'Euler basée sur la méthode des rectangles.

• $q = 2$ C'est l'exemple de la section précédente, basée sur la méthode des trapèzes, avec le tableau :

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \beta & \beta & 0 \\ \hline & 1 - \frac{1}{2\beta} & \frac{1}{2\beta} \end{array}$$

NB : $\beta = (2\alpha)^{-1}$.

• $q = 4$ : la méthode de Runge-Kutta "classique" (la plus utilisée) basée sur la formule de Simpson :

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 2/6 & 2/6 & 1/6 \end{array}$$

La méthode s'écrit aussi :

$$\Phi(t, y, h) = \frac{1}{6} \left[k_1 + 2k_2 + 2k_3 + k_4\right]$$

avec :

$$\begin{aligned} k_1 &= f(t,y) \\ k_2 &= f(t+\frac{h}{2}, y+\frac{h}{2}k_1) \\ k_3 &= f(t+\frac{h}{2}, y+\frac{h}{2}k_2) \\ k_4 &= f(t+h, y+hk_3) \end{aligned}$$

La méthode est d'ordre 4 mais il vaut mieux avoir Maple pour le vérifier !

## 2.5 À vous : quelques modèles à traiter

**Exercice 1.** *Décrire le mouvement d'une bille roulant sans frottement sur un plan incliné.*

**Exercice 2.** *Mouvements des planètes, équations de Képler*

*On considère une planète assimilée à un point matériel qui tourne autour du soleil (dans $\mathbb{R}^3$). Si x(t) désigne sa position à l'instant t, le soleil étant à l'origine, on a, d'après la loi de la gravitation universelle :*

$$\ddot{x}(t) = -\frac{c}{|x(t)|^3}x(t) \ ,$$

*où c est une constante qui prend en compte les différentes masses et la constante de gravitation g.*
*(i) Montrer que, si $\wedge$ est le produit vectoriel dans $\mathbb{R}^3$, $x(t)\wedge\dot{x}(t)$ est constant. En déduire que le mouvement est plan et suit la "loi des aires" (l'aire balayée par le vecteur $x(t)$ pendant l'intervalle de temps $[t_1, t_2]$ ne dépend que de $t_2 - t_1$.*
*(ii) Comme le mouvement est plan, on peut écrire $x(t) = r(t)\exp(i\theta(t))$. Donner les équations satisfaites par $r(t)$ et $\theta(t)$ et retrouver la loi des aires.*
*(iii) On suppose que $\theta$ est une fonction strictement croissante de t [on pourra se demander pourquoi on peut supposer cela]. Si $\psi(\theta) = t$ (au moins localement) et on s'intéresse à $u(\theta) = [r(\psi(\theta))]^{-1}$. Prouver que :*

$$u''(\theta) = -u(\theta) + l \ ,$$

*où l est une constante et en déduire que :*

$$r(\psi(\theta)) = \frac{p}{1 + e\cos(\theta)} \ ,$$

*et que la planète décrit une ellipse.*
*(iv) La loi de la gravitation universelle dérive-t-elle d'un potentiel ? Si oui, écrire l'énergie conservée et dire quand la vitesse de la planète est minimale et maximale sur sa trajectoire.*

**Exercice 3.** *Équation logistique :*
*Quand on s'intéresse à l'évolution d'une population de taille $N(t)$ à l'instant $t$, on doit prendre en compte un taux de natalité $n$ et de mortalité $d$ et l'équation la plus simple que l'on peut écrire est :*

$$\dot{N}(t) = nN(t) - dN(t) = aN(t) \ ,$$

*où $a = n - d$. Comme échauffement, on pourra calculer les solutions de cette équation pour $a$ constant ou dépendant du temps $t$ et interpréter les résultats.*
*Mais ces modèles sont trop naïfs et les biologistes préfèrent des lois "logistiques" :*

$$\dot{N}(t) = \gamma N(t)(N^* - N(t)) \ ,$$

*où $\gamma$ et $N^*$ sont des constantes positives. L'interprétation de cette équation, c'est que le milieu où vit cette population ne peut nourrir qu'un nombre limité d'individu (essentiellement $N^*$) et, arrivé à ce seuil, la population ne peut plus croître. Résoudre explicitement cette edo et interpréter les résultats.*
*Même question pour la dynamique de type Michaelis-Menten :*

$$\dot{N}(t) = \frac{aN(t)}{b + N(t)} \ ,$$

*où $a, b > 0$, équation utilisée en chimie et en biologie.*

**Exercice 4.** *Epidémiologie, évolution de population*

*La biologie fournit un nombre de modèles très variés : un des plus connus est le modèle proies-prédateurs qui conduit au système différentiel de Volterra-Lotka*

$$\frac{dL(t)}{dt} = aL - bLR \ ,$$
$$\frac{dR(t)}{dt} = -pR + qLR \ .$$

*où $L(t)$ est le nombre de proies à l'instant $t$, $R(t)$ le nombre de prédateurs et $a$, $b$, $p$, $q$ sont des constantes positives.*

*Les différents termes s'interprètent comme suit :*

*– $AL$ est la croissance du nombre des proies grâce aux naissances.*

*– $-bLR$ est sa diminution due aux repas des prédateurs !*

*– $-pR$ est la diminution du nombre de prédateurs en l'absence de proies.*

*– Dès qu'il y a présence de proies, le nombre de prédateurs augmente et ceci d'autant plus vite qu'il y a de proies d'où le terme $qLR$.*

*Il s'agit là encore d'un cas modèle car l'étude mathématique de ce système d'équations n'est pas trop difficile. De plus, on simule facilement (en utilisant par exemple la méthode de Runge-Kutta) les évolutions des populations régies par ces équations à partir d'un quelconque état initial.*

*Étudier ce système en montrant qu'il existe une intégrale première de la forme :*

$$c_1 L(t) + c_2 R(t) + c_3 \log(L(t)) + c_4 \log(R(t)) \ ,$$

*en caculant le ou les points d'équilibre et en discutant sa (ou leur) stabilité. Programmer un schéma d'Euler et un schémas RK4 : que constate-t-on ?*

# 3 Modélisation et équations aux dérivées partielles

## 3.1 Évolution de la température le long d'une barre : modélisation et premières remarques

On considère une barre infiniment longue et on note $x$ la coordonnée le long de la barre. A l'instant $t = 0$, on observe une température $T_0(x)$ au point $x$. Cette chaleur va se diffuser le long de la barre et on voudrait connaître son évolution au cours du temps.

On note $T(x,t)$ la température au point $x$ à l'intant $t$. La variation de la quantité de chaleur dans un intervalle $[x - \Delta x, x + \Delta x]$ est égale au flux de chaleur à travers les bord (les points $x - \Delta x, x + \Delta x$) et la loi de Fourier indique que ce flux de chaleur est proportionnel au gradient de $T$. D'où :

$$\frac{d}{dt}\left[\int_{x-\Delta x}^{x+\Delta x} T(y,t)dy\right] = -\lambda \frac{\partial T}{\partial x}(x - \Delta x, t) + \lambda \frac{\partial T}{\partial x}(x + \Delta x, t) \ ,$$

où $\lambda > 0$ est la conductivité thermique du milieu. Le changement de signe s'explique par le changement de signe du flux. On en déduit :

$$\int_{x-\Delta x}^{x+\Delta x} \frac{\partial T}{\partial t} dy = \int_{x-\Delta x}^{x+\Delta x} \lambda \frac{\partial^2 T}{\partial x^2}(y,t)dt \ ,$$

Comme $\Delta x$ est arbitraire, la fonction $T$ satisfait l'équation de la chaleur :

$$\frac{\partial T}{\partial t} - \lambda \frac{\partial^2 T}{\partial x^2} = 0 \quad \text{dans } \mathbb{R} \times (0, t_{max}) \ . \tag{1}$$

(on prendra désormais $\lambda = 1/2$ pour simplifier).

La première remarque c'est que la modélisation s'appuie effectivement ici sur des lois de la physique qui existent et sont bien établies.

L'équation ci-dessus revient à peu près à dire que si on connait la température $T$ à l'instant $t$, on la calcule à l'instant $t + \Delta t$ en moyennant la température autour de $x$ de la manière suivante :

$$T(x, t + \Delta t) \simeq \frac{1}{2}\left[T(x + \sqrt{\Delta t}, t) + T(x - \sqrt{\Delta t}, t)\right] \ . \tag{2}$$

Ce cas est un exemple modèle car :

1. On connait explicitement la solution :

$$T(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} T_0(y) \exp\left(-\frac{|x-y|^2}{2t}\right) dy \ . \tag{3}$$

   Et, grâce à cela, on a une étude mathématique complète. Rarissime !
2. Du point de vue physique, ce modèle a un problème majeur : les informations se propagent à vitesse infinie ce qui n'est pas physiquement acceptable. En effet, si $T_0 > 0$ sur un petit intervalle centré en 0 et nul ailleurs, alors $T(x,t) > 0$ pour tout temps $t > 0$ et pour tout $x$. On n'a pas ici de propriété de vitesse finie de propagation : c'est une différence fondamentale entre équations paraboliques et équations hyperboliques que nous verrons par la suite.
3. Pourtant, la confrontation avec l'expérience est très satisfaisante : ce qui prouve que la modélisation même très simplifiée peut donner de bons résultats en dépit de problèmes apparents. Et justement l'avantage de ce modèle, c'est sa simplicité.
4. Du point de vue simulation, on ne calcule JAMAIS les valeurs de la fonction $T$ avec la formule explicite (3) (sauf évidemment quand l'intégrale se calcule explicitement...) mais plutôt avec des schémas numériques de type (2), en général plus efficaces que celui-là.

## 3.2 Évolution de la température le long d'une barre : étude mathématique

On commence par le :

**Théorème 4.** *Si $T_0$ est une fonction continue qui satisfait la propriété :*

$$\lim_{|x|\to+\infty} T_0(x)\varepsilon^{-A|x|^2} = 0 \tag{4}$$

*pour une certaine constante $A > 0$, alors la fonction $T$ définie par (3) est une solution de (1) avec la donnée initiale :*

$$T(x,0) = T_0(x) \quad \textit{dans } \mathbb{R} \ , \tag{5}$$

*et cette solution est définie sur un intervalle de temps $[0, t_{max}[$ avec $t_{max} = 1/(2A)$. De plus, cette solution est unique dans la classe des solutions qui croissent au plus en $\exp(Cx^2)$ à l'infini.*

**Remarque :** si $t \in ]0, t_{max}[$ avec $t_{max} = 1/(2A)$, l'intégrale écrite dans la proposition a bien un sens, d'où le choix de $t_{max}$ qui est optimal.

**Preuve :** On va calculer la solution fondamentale de l'équation (1) *i.e.* la solution $\rho$ de (1) associée à la donnée initiale suivante :

$$\rho(x,0) = \delta_0(x) \quad \text{dans } \mathbb{R} \, , \tag{6}$$

où $\delta_0$ est la masse de Dirac centrée en $x = 0$. On notera $\rho(x,t)$ cette solution. La solution $T$ sera alors donnée à partir de $\rho$ via la formule :

$$T(x,t) = \Big(\rho(\cdot,t) * T_0\Big)(x)$$

où la convolution agit dans la variable d'espace seulement. En effet, d'après les propriétés de la convolution, si $\rho$ est une fonction régulière,

$$\frac{\partial T}{\partial t} - \frac{1}{2}\frac{\partial^2 T}{\partial x^2} = \Big(\frac{\partial \rho}{\partial t} - \frac{1}{2}\frac{\partial^2 \rho}{\partial x^2}\Big) * T_0 = 0$$

et :

$$\Big(\rho(\cdot,t) * T_0\Big)(x) \to T_0(x) \quad \text{quand} \quad t \to 0$$

(ces propriétés deviendront plus claires après le calcul explicite de $\rho$).

Maintenant, pour calculer $\rho$, on raisonne formellement en introduisant la transformée de Fourier en $x$ :

$$\hat{\rho}(\xi,t) = \int_{\mathbb{R}} \rho(x,t) e^{-\mathrm{i}\xi x}\, \mathrm{d}x$$

et on applique cette transformée de Fourier à l'équation. Il est bien connu que :

$$\begin{aligned} \widehat{\frac{\partial \rho}{\partial t}}(\xi,t) &= \frac{\partial \hat{\rho}}{\partial t}(\xi,t) \\ \widehat{\frac{\partial^2 \rho}{\partial x^2}}(\xi,t) &= -\xi^2 \hat{\rho}(\xi,t) \end{aligned}$$

d'où :

$$\Big(\frac{\partial \hat{\rho}}{\partial t} + \xi^2 \hat{\rho}\Big)(\xi,t) = 0 \quad \text{dans} \quad \mathbb{R} \times (0, t_{max})$$

avec :

$$\hat{\rho}(\xi,0) = 1 \quad \text{dans} \quad \mathbb{R}\,.$$

Il en résulte de l'intégration de cette équation différentielle en $t$ que :

$$\hat{\rho}(\xi,t) = \exp\Big(-\frac{t\xi^2}{2}\Big) \quad \text{dans} \quad \mathbb{R} \times (0, t_{max})\,.$$

Or la transformée de Fourier de $\exp(-a\xi^2)$ est : $\sqrt{\dfrac{\pi}{a}} \exp\big(-\dfrac{x^2}{4a}\big)$, de sorte que l'on obtient une formule explicite pour $\rho$ :

$$\rho(x,t) = \frac{1}{\sqrt{2\pi t}} \exp\Big(-\frac{x^2}{2t}\Big)\,,$$

puisqu'en effet la transformée de Fourier inverse est :

$$\rho(x,t) = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{\rho}(\xi,t) \varepsilon^{\mathrm{i}\xi x}\, \mathrm{d}\xi\,.$$

On peut aussi vérifier directement et de manière élémentaire (mais cela demande un peu de travail autour du Théorème de Lebesgue de dérivation sous le signe somme) que la fonction $T$ donnée dans l'énoncé de la proposition est bien solution de l'équation (exo !).

Pour l'unicité, on va prouver que l'on a une propriété un peu plus forte : le *Principe du Maximum.*

**Lemme 2.** *Si $T_1$ et $T_2$ sont des fonctions régulières ($C^2$ en $x$ et $C^1$ en $t$) sur $\mathbb{R} \times [0, t_{max}[$ qui satisfont la condition de croissance :*

$$|T_1(x,t)|, |T_2(x,t)| \leq C \exp(Cx^2) \quad \textit{dans } \mathbb{R} \times (0, t_{max}) \ ,$$

*pour une certainne constante $C > 0$ et :*

$$\frac{\partial T_1}{\partial t} - \frac{1}{2}\frac{\partial^2 T_1}{\partial x^2} \leq 0 \quad \textit{dans } \mathbb{R} \times (0, t_{max}) \textit{ ($T_1$ est sous-solution de l'équation)},$$

$$\frac{\partial T_2}{\partial t} - \frac{1}{2}\frac{\partial^2 T_2}{\partial x^2} \geq 0 \quad \textit{dans } \mathbb{R} \times (0, t_{max}) \textit{ ($T_2$ est sursolution de l'équation)},$$

$$T_1(x,0) \leq T_2(x,0) \quad \textit{dans } \mathbb{R} \ ,$$

*alors :*

$$T_1(x,t) \leq T_2(x,t) \quad \textit{dans } \mathbb{R} \times (0, t_{max}) \ .$$

Évidemment ce résultat donne l'unicicité car si $T_1$ et $T_2$ sont deux solutions qui satisfont la condition de croissance alors on peut faire jouer à $T_1$ le rôle de sous-solution et à $T_2$ le rôle de sursolution et on conclut que $T_1 \leq T_2$ dans $\mathbb{R} \times (0, t_{max})$. Mais, en inversant les rôles, on a l'inégalité opposée donc l'égalité.

Pour *prouver le lemme*, on note $T = T_1 - T_2$ qui satisfait :

$$|T(x,t)| \leq 2C \exp(Cx^2) \quad \text{dans } \mathbb{R} \times (0, t_{max}) \ ,$$

pour une certainne constante $C > 0$ et :

$$\frac{\partial T}{\partial t} - \frac{1}{2}\frac{\partial^2 T}{\partial x^2} \leq 0 \quad \text{dans } \mathbb{R} \times (0, t_{max}) \ ,$$

$$T(x,0) \leq 0 \quad \text{dans } \mathbb{R} \ .$$

On considère maintenant :

$$\chi(x,t) := \exp((K_1 t + K_2) C x^2) \ ,$$

où $K_1, K_2 > 1$ sont des constantes à fixer. Des calculs élémentaires montrent que, si on choisit $K_2 > 1$ quelconque et $K_1$ assez grand alors :

$$\frac{\partial \chi}{\partial t} - \frac{1}{2}\frac{\partial^2 \chi}{\partial x^2} > 0 \quad \text{dans } \mathbb{R} \times (0, \tau] \ ,$$

pour $\tau$ assez petit (typiquement $K_1\tau = K_2$).

On raisonne alors de la manière suivante : pour $0 < \alpha \ll 1$, on considère :

$$\sup_{\mathbb{R}\times[0,\tau]} (T(x,t) - \alpha\chi(x,t)) \ .$$

Ce "sup" est en fait un "max" car $T(x,t)-\alpha\chi(x,t) \to -\infty$ quand $|x| \to +\infty$. Il existe donc $(x_0,t_0) \in \mathbb{R}\times[0,\tau]$ qui est un point de maximum de $T-\chi$ sur $\mathbb{R}\times[0,\tau]$.

Se présente alors deux cas :

– $t_0 = 0$ alors le maximum est négatifs puisque $T(x,0) \leq 0$ pour tout $x$ et $\chi \geq 0$. On a donc, si tel est le cas, $T(x,t) - \alpha\chi(x,t) \leq 0$ dans $\mathbb{R}\times[0,\tau]$.

– $t_0 > 0$ : comme on est en un point de maximum,

$$\frac{\partial T}{\partial t}(x_0,t_0) = \alpha\frac{\partial \chi}{\partial t}(x_0,t_0) \quad \text{et} \quad \frac{\partial^2 T}{\partial x^2}(x_0,t_0) \leq \alpha\frac{\partial^2 \chi}{\partial x^2}(x_0,t_0) \ ,$$

mais on aurait alors :

$$0 \geq \frac{\partial T}{\partial t}(x_0,t_0) - \frac{1}{2}\frac{\partial^2 T}{\partial x^2}(x_0,t_0) \geq \alpha(\frac{\partial \chi}{\partial t}(x_0,t_0) - \frac{\partial^2 \chi}{\partial x^2}(x_0,t_0)) > 0 \ ,$$

une contradiction.

On est donc toujours dans le premier cas, $T(x,t) - \alpha\chi(x,t) \leq 0$ dans $\mathbb{R}\times[0,\tau]$ pour tout $\alpha > 0$ et on conclut en faisant tendre $\alpha$ vers 0.
**NB :** une petite erreur s'est glissée dans le raisonnement ci-dessus : la trouver ! □

**Exercice 5. L'équation de Black & Scholes en Finance :**
*Elle s'écrit :*

$$-\frac{\partial U}{\partial t} + rU - rP\frac{\partial U}{\partial P} - \frac{1}{2}\sigma^2 P^2\frac{\partial^2 U}{\partial P^2} = 0 \quad \textit{pour } P > 0, t \leq T \ ,$$

*où $U$ est le prix d'une* option d'achat *qui rapporte $(P-K)^+$ si on attend la maturité $T$ [on a donc $U(P,T) = (P-K)^+$], $P$ désigne le prix d'une action, $\sigma$ sa volatilité (taux de fluctuation aléatoire de l'action), $r$ est le taux court (taux de caisse d'épargne). Le but de l'exercice est de calculer $U$ "explicitement" en se ramenant à (3) via divers changement de variables. On commencera par poser $P = \exp(x)$ et :*

$$u(x,t) = U(P, T-t) \ ,$$

*puis :*

$$v(x,t) = \exp(Kt)u(x-ct,t) \ .$$

## 3.3 Évolution de la température le long d'une barre : étude numérique

On va s'intéresser à la résolution numérique de :

$$\frac{\partial u}{\partial t} - \frac{1}{2}\frac{\partial^2 u}{\partial x^2} = 0 \quad \text{dans } \mathbb{R} \times (0,T)\,. \tag{7}$$

$$u(x,0) = u_0(x) \quad \text{dans } \mathbb{R}\,. \tag{8}$$

Dans un premier temps, on ne va pas se soucier d'une difficulté essentielle de la résolution numérique de (7)-(8) : le fait que l'équation soit posée dans un domaine non borné...

Pour calculer numériquement la solution de (7)-(8), on introduit une grille en $x$ et en $t$, en posant :

$$\begin{aligned} x_j &= j\Delta x, \quad j \in \mathbb{Z}, \\ t_n &= n\Delta t, \quad n \in \{0,1,\cdots,N\},\ N\Delta t = T, \end{aligned}$$

et on notera $u_j^n$ une approximation de $u(x_j,t_n)$. Il est à remarquer cette fois que $j \in \mathbb{Z}$, ce qui n'est évidemment pas très compatible avec un calcul sur ordinateur ! Pour simplifier la présentation, on va donc d'abord présenter des schémas numériques définis pour $j \in \mathbb{Z}$ puis on expliquera comment se ramener à un nombre fini de valeurs de $j$.

## 3.4 Mise en place des principaux schémas

Il y a trois schémas numériques "classiques" pour calculer la solution de (7)-(8) qui sont les suivants : le schéma explicite standard (SES), le schéma implicite standard (SIS) et les $\theta$-schémas de Crank-Nickolson ; nous n'étudierons ici en détail que les deux premiers.

Schéma Explicite Standard :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{1}{2}\frac{u_{j+1}^n + u_{j-1}^n - 2u_j^n}{(\Delta x)^2} = 0\,,\ n \in \mathbb{N},\ j \in \mathbb{Z}\,. \tag{SES}$$

Comme dans le cas des équations hyperboliques (déjà vu ?), on étudie la stabilité par une analyse de Fourier ; si $u_j^n = \exp(\mathrm{i}kj\Delta x)$, alors :

$$u_j^{n+1} = g(\lambda, k\Delta x)u_j^n, \quad \text{avec} \quad \lambda = \frac{\Delta t}{2(\Delta x)^2}\,.$$

Le calcul de $g$ donne :

$$\begin{aligned} g(\lambda, k\Delta x) &= 1 + 2\lambda\big(\cos(k\Delta x) - 1\big) \\ &= 1 - 4\lambda \sin^2\Big(\frac{k\Delta x}{2}\Big)\,. \end{aligned}$$

Tout d'abord, comme $\lambda \geq 0$, on a bien $g(\lambda, k\Delta x) \leq 1$. Pour avoir maintenant $g(\lambda, k\Delta x) \geq -1$, il est nécessaire d'avoir $\lambda \leq 1/2$, *i.e.*

$$\frac{\Delta t}{(\Delta x)^2} \leq 1\,,$$

c'est la condition de CFL. Il est à noter que cette condition est équivalente à la monotonie du schéma ; en effet, on peut le réécrire sous la forme suivante :

$$u_j^{n+1} = \lambda u_{j+1}^n + \lambda u_{j-1}^n + (1-2\lambda)u_j^n$$

de telle sorte que, si $\lambda \leq 1/2$, on a la propriété de monotonie :

$$u_j^n \geq 0 \text{ pour tout } j \in \mathbb{Z} \Longrightarrow u_j^{n+1} \geq 0 \text{ pour tout } j \in \mathbb{Z}\,,$$

ou encore :

$$u_j^n \geq v_j^n \;\text{ pour tout }\; j \in \mathbb{Z} \Longrightarrow u_j^{n+1} \geq v_j^{n+1} \;\text{ pour tout }\; j \in \mathbb{Z}\,.$$

On retrouve ainsi une version "discrète" de la propriété de principe de maximum (et de comparaison) satisfaite par l'équation.

Le schéma (SES) est consistant, un calcul simple montre qu'il est d'ordre **2** en $x$ et d'ordre **1** en $t$.

**Remarque :** La condition de CFL est désagréable car elle implique que l'on doit prendre $\Delta t$ de l'ordre de $(\Delta x)^2$, c'est-à-dire extrêmement petit. On "avance" ainsi très peu vite en temps et pour calculer $u(x,t)$ il faut beaucoup d'itérations. En revanche, à chaque pas d'itération, le schéma calcule vite la solution puisqu'il est explicite [(2)].

Schéma Implicite Standard :

$$\text{(SIS)} \qquad \frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{1}{2}\frac{u_{j+1}^{n+1} + u_{j-1}^{n+1} - 2u_j^{n+1}}{(\Delta x)^2} = 0\,,\ n \in \mathbb{N},\ j \in \mathbb{Z}\,.$$

Pour étudier la stabilité de ce schéma, on peut adopter deux stratégies : une par Fourier, l'autre par principe de maximum.

Par Fourier, on a :

$$g(\lambda, k\Delta x) = \Big[1 - 2\lambda\big(\cos(k\Delta x) - 1\big)\Big]^{-1} \leq 1\,,$$

car $\cos(k\Delta x) - 1 \leq 0$, donc le schéma implicite est inconditionnellement stable.

(2). ceci est à comparer avec le schéma implicite standard pour lequel on est pas obligé de prendre $\Delta t$ petit, mais en revanche il faut résoudre un système linéaire à chaque itération - le choix d'un "bon" schéma n'est donc pas évident *a priori*.

Cette propriété peut être obtenue par un argument heuristique de type principe de maximum. Supposons $(u_j^n)_{j\in\mathbb{Z}}$ borné et montrons que :

$$\max_{j\in\mathbb{Z}} |u_j^{n+1}| \leq \max_{j\in\mathbb{Z}} |u_j^n| \,.$$

On ne va démontrer que l'inégalité :

$$\max_{j\in\mathbb{Z}} u_j^{n+1} \leq \max_{j\in\mathbb{Z}} u_j^n \,,$$

celle avec les valeurs absolues s'obtenant ou bien en raisonnant de la même façon avec le min, ou bien en changeant les $u_j^{n+1}, u_j^n$ en $-u_j^{n+1}, -u_j^n$ et en utilisant le résultat pour le max pour les $-u_j^{n+1}, -u_j^n$.

Supposons donc que le premier maximum est atteint en $(n+1, j_0)$. Cela entraine, en utilisant le schéma :

$$u_{j_0}^{n+1} - \frac{1}{2}\frac{\Delta t}{(\Delta x)^2}\big(u_{j_0+1}^{n+1} + u_{j_0-1}^{n+1} - 2u_{j_0}^{n+1}\big) = u_{j_0}^n \,.$$

Mais $u_{j_0+1}^{n+1} \leq u_{j_0}^{n+1}$ et $u_{j_0-1}^{n+1} \leq u_{j_0}^{n+1}$, donc :

$$u_{j_0+1}^{n+1} + u_{j_0-1}^{n+1} - 2u_{j_0}^{n+1} \leq 0 \,,$$

(ce qui n'est rien d'autre qu'une version "discrète" du fait que la dérivée seconde est négative au sens large en un point de maximum !) et ainsi :

$$\max_{j\in\mathbb{Z}} u_j^{n+1} = u_{j_0}^{n+1} \leq u_{j_0}^n \leq \max_{j\in\mathbb{Z}} u_j^n \,.$$

Il est à noter que cette propriété signifie que le schéma est inconditionnellement monotone, *i.e.*

$$u_j^n \leq 0 \;\text{ pour tout } j\in\mathbb{Z} \Longrightarrow u_j^{n+1} \leq 0 \quad \text{pour tout } j\in\mathbb{Z} \,.$$

Comme le schéma (SES), le schéma (SIS) est consistant et d'ordre **1** en temps, **2** en espace.

### 3.5 Convergence des schémas

On a le résultat suivant :

**Théorème 5.** *On suppose que $u_0$ est bornée, uniformément continue sur $\mathbb{R}$. Alors la solution $(u_j^n)_{n,j}$ des schémas (SES) quand la condition de CFL est satisfaite et (SIS) (inconditionnellement) converge uniformément vers $u$ sur $\mathbb{R}\times[0,T]$ pour tout $T>0$, i.e.*

$$\max_{\substack{j\in\mathbb{Z}\\ n\in\mathbb{N}}} \big|u_j^n - u(j\Delta x, n\Delta t)\big| \to 0 \quad \textit{quand} \quad \Delta x + \Delta t \to 0 \,.$$

**Preuve :**
ETAPE 1 : on traite le cas où $u_0$ est de classe $C^4$ avec $u_0^{(k)}$ bornées pour $0 \leq k \leq 4$. On injecte les valeurs de $u(j\Delta x, n\Delta t)$ dans le schéma et on utilise la monotonie pour montrer que :

$$\left|u_j^n - u(j\Delta x, n\Delta t)\right| \leq Cn\Delta t\left[(\Delta x)^2 + \Delta t\right],$$

où $C[(\Delta x)^2 + \Delta t]$ est le terme d'erreur que l'on commet à chaque étape en remplaçant $u_j^n$ par $u(j\Delta x, n\Delta t)$, et :

$$Cn\Delta t\left[(\Delta x)^2 + \Delta t\right]$$

et une "sur-solution" qui permet de prendre en compte cette erreur.
ETAPE 2 : on approche $u_0$ qui est bornée et uniformément continue par des fonctions $(u_0^\varepsilon)_\varepsilon$ qui sont de classe $C^4$ avec les quatre permières dérivées bornées. Cette approximation uniforme (obtenue par convolution) est rendue possible par l'uniforme continuité de $u_0$. On raisonne ensuite comme dans le Chapitre 1, en utilisant le formule explicite qui donne $u, u^\varepsilon$,... etc. □

**Remarque :** Dans le cas où $u_0$ est $C^4$, avec $u_0^{(k)}$ borné pour $0 \leq k \leq 4$, on a bien une convergence en $(\Delta t)^{\mathbf{1}} + (\Delta x)^{\mathbf{2}}$ comme le prévoit l'ordre du schéma.

### 3.6 Aproximation par domaine de calcul borné

A cause de la vitesse infinie de propagation des informations, ce problème apparait comme non trivial. Pourtant, si l'on considère par exemple le problème suivant :

$$\text{(1')} \qquad \frac{\partial u^R}{\partial t} - \frac{1}{2}\frac{\partial^2 u^R}{\partial x^2} = 0 \quad \text{dans } \ ]-R, R[\times(0,\infty)$$

avec la condition initiale suivante :

$$\text{(5')} \qquad u^R(x,0) = u_0(x) \quad \text{dans } \ ]-R, R[$$

et la condition de bord :

$$\text{(4')} \qquad u^R(\pm R, t) = u_0(\pm R) \quad \text{pour} \quad t > 0\,,$$

on peut montrer que $u^R \to u$ localement uniformément quand $R \to \infty$. La semi-discrétisation en temps que nous avons évoqué montre que ce problème se ramène à ce que nous avons vu au Chapitre 1.
**N.B. :** la condition aux limites artificielle (4') peut (en fait, doit !) être remplacée la plupart du temps par une condition plus adaptée et plus astucieuse ! ! !

**Exercice 6.** *Programmer au moins l'un des schémas et examiner l'effet des conditions aux limites artificielles.*

## 3.7 Équation de la chaleur en domaines bornés

On peut s'intéresser à l'équation de la chaleur dans un domaine borné (par exemple $[0,1]$) :

$$\frac{\partial u}{\partial t} - \frac{1}{2}\frac{\partial^2 u}{\partial x^2} = 0 \quad \text{dans } [0,1]\times(0,T)\ . \tag{9}$$

avec une donnée initiale :

$$u(x,0) = u_0(x) \quad \text{dans } [0,1]\ . \tag{10}$$

Mais il faut ajouter des conditions aux limites sur les bords de l'intervalle $[0,1]$ (ou sur les ensembles $\{0\}\times(0,T)$ et $\{1\}\times(0,T)$). Parmi les diverses conditions existentes, deux jouent un rôle plus importants :

1. Conditions de Dirichlet : $u(a,t) = g(t)$ pour $a = 0$ et/ou 1. On fixe donc la température à une extrémité de l'intervalle ou aux deux.
2. Conditions de Neumann : $\dfrac{\partial u}{\partial x} = h(t)$. Ici on fixe le flux de chaleur. Par exemple, si $h \equiv 0$, on a une condition d'adiabaticité : aucun flux de chaleur ne traverse la paroi (isolation parfaite).

Évidemment on peut combiner ces deux types de conditions au limites en imposant Dirichlet à une extremité et Neumann à l'autre.

Pour traiter ce problème, nous allons disposer d'un outil différent de type série de Fourier. Pour justifier son introduction, on va considérer des conditions de Dirichlet homogène :

$$u(0,t) = u(1,t) = 0 \quad \text{pour tous } t > 0\ , \tag{11}$$

et on va chercher des solutions à variables séparées donc de la forme :

$$u(x,t) = \psi(t)\phi(x)\ .$$

On a donc :

$$\psi'(t)\phi(x) - \psi(t)\phi''(x) = 0\ ,$$

pour tous $x$ et $t$. Seule possibilité pour obtenir des solutions non triviales :

$$\psi'(t) = \lambda\psi(t) \text{ pour tous } t > 0,$$

$$\lambda\phi(x) - \phi''(x) = 0 \quad \text{dans } (0,1)\ ,$$

sans oublier la condition aux limites :

$$\phi(0) = \phi(1) = 0\ .$$

Pour $\psi$, on a $\psi(t) = c\exp(\lambda t)$ où $c \in \mathbb{R}$ et pour $\phi$, on a un problème de type valeur propre qui conduit à :

$$\lambda_k = k^2\pi^2 \quad \text{et} \quad \phi_k(x) = \sin(k\pi x) \ ,$$

pour $k \in \mathbb{N}^*$.

Phrase compliquée : la théorie des opérateurs compacts dans $L^2([0,1])$ implique que les $\phi_k$ forment une base hilbertienne de $L^2([0,1])$ (l'opérateur compact est $T : L^2([0,1]) \to L^2([0,1])$ tel que $T(f) = u$ où $u$ est l'unique solution de :

$$-u''(x) = f \quad \text{dans } (0,1) \ ,$$

avec la condition aux limites :

$$u(0) = u(1) = 0 \ .)$$

Quoiqu'il en soit, si on peut écrire que :

$$u_0(x) = \sum_{k=1}^{+\infty} a_k \phi_k(x) \ ,$$

alors la solution est donnée par :

$$u(x,t) = \sum_{k=1}^{+\infty} a_k \exp(-k^2\pi^2 t)\phi_k(x) \ ,$$

et on voit facilement que même si $u_0 \in L^2([0,1])$ alors $u$ est $C^\infty$ (car ? ? ? ?).

**Exercice 7.** *Faire le même travail pour l'équation des cordes vibrantes (fixées aux extrémités) :*

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0 \quad \textit{dans } \mathbb{R} \times (0,T) \ , \tag{12}$$

*où $c > 0$. Quelles sont les différences avec l'équation de la chaleur ? Comment s'interprète c ?*

## 3.8 Quelques éléments de modélisation du trafic routier

On va reprendre quelques éléments de la modélisation conduisant à l'équation de la chaleur mais dans un contexte différent.

Cette fois la droite réelle est une route à sens unique (de la gauche vers la droite) où se déplacent des véhicules.

On note $u(x,t)$ la densité de véhicules au point $x$ à l'intant $t$. La variation de la quantité de véhicules dans un intervalle $[x - \Delta x, x + \Delta x]$ est égale au

flux entrant (par le point $x - \Delta x$) et sortant (par le point $x + \Delta x$). Si $\phi$ est ce flux, on a :

$$\frac{d}{dt}\left[\int_{x-\Delta x}^{x+\Delta x} u(y,t)dy\right] = \phi(x-\Delta x,t) - \phi(x+\Delta x,t) \ ,$$

Le changement de signe du flux est ici évident. On en déduit :

$$\int_{x-\Delta x}^{x+\Delta x} \frac{\partial u}{\partial t} dy = \int_{x-\Delta x}^{x+\Delta x} -\frac{\partial \phi(x,t)}{\partial x}(y,t)dt \ ,$$

et comme $\Delta x$ est arbitraire, les fonctions $u$ et $\phi$ sont liées par l'équation :

$$\frac{\partial u}{\partial t} + \frac{\partial \phi}{\partial x} = 0 \quad \text{dans } \mathbb{R} \times (0, +\infty) \ . \tag{13}$$

Cette équation ayant deux fonctions inconnues, il n'est pas possible de la résoudre sans lier ces deux fonctions par une *loi de comportement* liant le flux de véhicule et sa densité. Si on normalise $u$ de telle sorte que $0 \leq u \leq 1$ alors on peut penser a une loi du type :

$$\phi = \frac{1}{2}u(1-u) \ .$$

Le flux est petit si $u$ est proche de 0 car il y a peu de véhicules, il est petit quand il y a trop de véhicules (i.e. pour $u$ proche de 1) à cause de l'engorgement et $u = 1/2$ donne le flux maximal (trafic ni trop important, ni trop faible).

En posant $v = 1/2 - u$, comme $\phi = -v^2/2 + 1/8$, on se ramène à l'équation de Burgers :

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = 0 \quad \text{dans } \mathbb{R} \times (0, +\infty) \ . \tag{14}$$

Le cours sur les équation de transport indique que, pour résoudre de telles équations, on doit chercher des courbes caractéristiques $(x(t), t)$ le long desquelles $v$ est constante. En dérivant $v(x(t), t) = \textit{constante}$, il vient :

$$\frac{\partial v}{\partial t} + \dot{x}(t)\frac{\partial v}{\partial x} = 0$$

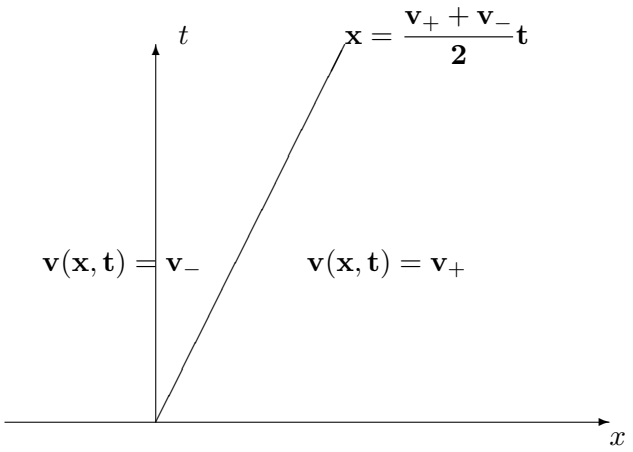
et en identifiant avec l'équation, on voit que $\dot{x}(t) = v(x(t), t) = v(x(0), 0)$. Les courbes caractéristiques sont des droites et l'on voit facilement que ces droites peuvent se couper : ceci signifie que les solutions de ces équations ne sont pas régulières en général (sinon le calculs ci-dessus serait justifié et on aurait une contradiction...) et il faut employer la notion de solution au sens des distributions (ce qui ne suffit toujours pas...). Il y a, en général, plusieurs solutions au sens des distributions et il faut employer la notion de *solutions entropiques* qui ne rentre pas dans le cadre de ce cours.

Le *problème de Riemann* est un cas particulièrement intéressant car on sait calculer l'unique solution entropique pour l'équation de Burgers. C'est le cas où la donnée initiale est donnée par :

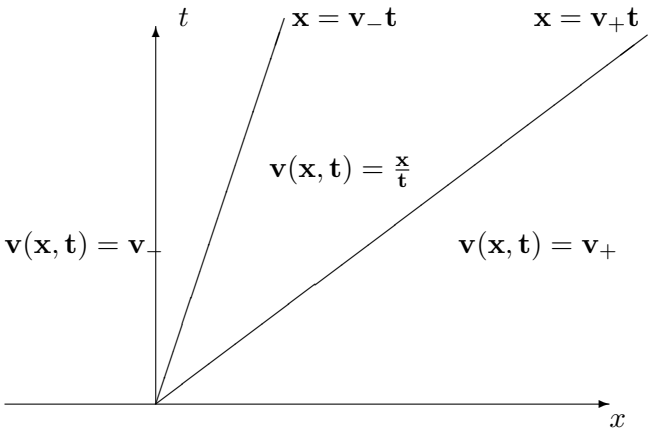
$$v(x,0) = \begin{cases} v_- & \text{si } x < 0 \ , \\ v^+ & \text{si } x \geq 0 \ . \end{cases}$$

On a deux cas :

**Si $\mathbf{v_- > v_+}$**, la solution présente une discontinuité et elle est donnée par :




**Si $\mathbf{v_- < v_+}$**, la solution est continue et elle est donnée par :




Dans le cas du trafic routier, on a une interprétation assez simple de ces solutions. Rappelons d'abord que la densité de véhicule $u$ vaut $1 - v$.

Si $v_- > v_+$, comme le trafic a lieu vers la droite, il y a plus de véhicule en aval qu'en amont : les véhicules arrivant de l'amont se retrouve donc face à une concentration de véhicules plus grande devant eux et donc face à un embouteillage (un "choc"). Au contraire, $v_- < v_+$, il y a plus de véhicule en amont qu'en aval : les véhicules arrivant de l'amont se retrouve donc face à une concentration de véhicules plus faible devant eux. La route se dégage, c'est une "onde de détente".

**Exercice 8.** *Vérifier que la première solution est TOUJOURS solution au sens des distributions (que l'on ait $v_- > v_+$ ou $v_- < v_+$) et que la deuxième l'est dans le cas où elle est bien définie, i.e. $v_- < v_+$.*

Reste à calculer numériquement ces solutions : les trois schémas les plus classiques se mettent sous forme conservative :

$$v_j^{n+1} = v_j^n - \lambda \left( g(v_j^n, v_{j+1}^n) - g(v_{j-1}^n, v_j^n) \right) ,$$

où $g$ est une fonction telle que $g(v, v) = v^2/2$ et $\lambda = \Delta t/\Delta x$.

Le schéma le plus spécifique dans le cas non linéaire est celui de *Godounov* où :

$$g(v_-, v_+) = w(0, v_-, v_+) ,$$

où $w(0, v_-, v_+)$ est la solution du problème de Riemann associée aux données $v_-, v_+$ qui ne dépend pas de $t$ vu la forme des solutions ($w = v_+, v_-, 0$ suivant divers cas).

Le schémas de *Lax-Friedrichs* est celui où :

$$g(v_-, v_+) = \frac{1}{2} \left( \frac{1}{2}(v_+^2 + v_-^2) - \lambda^{-1}(v_+ - v_-) \right) ,$$

et celui de *Lax-Wendroff* est donné par :

$$g(v_-, v_+) = \frac{1}{2} \left( \frac{1}{2}(v_+^2 + v_-^2) - \lambda^{-1}(\frac{v_+ + v_-}{2})(v_+^2 - v_-^2) \right) .$$

**Exercice 9.** *Vérifier la consistance de ces schémas, programmer les et comparer les.*

# 4 Modélisation et optimisation

On se propose ici de traiter un problème particulier sur la composition chimique d'un mélange de gaz à pression et température données (texte proposé à l'épreuve de Calcul scientifique de l'agrégation).

Après réaction chimique, un mélange de différents gaz à température $T$ et pression $p$ fixées doit normalement se trouver à l'équilibre thermodynamique.

On suppose que ce mélange est composé de $n_s$ espèces chimiques contenant au total $n_e$ types d'atomes différents, que l'on nomme les éléments.

Une mole de l'espèce $k$ est la quantité de matière correspondant à un nombre de molécules de cette espèce égal au nombre d'Avogadro. On note $N_k$ le nombre de moles de l'espèce $k$ et on introduit le vecteur $N = (N_1, \cdots, N_{n_s})$ de $\mathbb{R}^{n_s}$ et si $\bar{N} = N_1 + \cdots + N_{n_s}$, on appelle fraction molaire de l'espèce $k$, la quantité :

$$X_k = \frac{N_k}{\bar{N}}.$$

Et $X = (X_1, \cdots, X_{n_s})$. À l'équilibre chimique, les $N_k$ et les $X_k$ sont des inconnues.

Le nombre d'atomes de l'élément numéro $j$ dans une molécule de l'espèce $k$ sera noté $E_{kj}$ et on note $E$ la matrice $(E_{kj})_{kj}$. C'est une matrice dont les éléments sont des entiers positifs ou nuls donnés. Chaque ligne et chaque colonne de E admet au moins un élément non nul.

Le nombre de moles d'atome $j$ dans le mélange est noté $N_j^e$ et on a :

$$N_j^e = \sum_{k=1}^{n_s} N_k.E_{kj} \ .$$

On regroupe également ces valeurs en un vecteur $N^e$

On considère un système fermé, donc le nombre de moles de chaque type d'atomes se conserve au cours de la réaction. Ceci signifie que le vecteur $N^e$ est une donnée du problème.

Le vecteur des nombres de moles $N$ est donc soumis à $n_e$ contraintes d'égalité, linéaires, indépendantes si l'on fait l'hypothèse que la matrice E est de rang $n_e$ et qui s'écrivent :

$$N^e = N.E = N_0.E \ , \tag{15}$$

si $N_0$ est la composition initiale du mélange de gaz (avant les réactions conduisant à l'équilibre thermodynamique).

À $p$ et $T$ fixées, une composition chimique d'équilibre, c'est-à-dire le vecteur $N$ que l'on cherche à déterminer, est celle qui réalise un minimum d'une fonction G (appelée enthalpie libre ou énergie de Gibbs), sous les contraintes ci-dessus. Ici, G peut s'expliciter sous la forme :

$$G(N) = \sum_{k=1}^{n_s} N_k(g_k + \log(X_k)) = \sum_{k=1}^{n_s} N_k(g_k + \log(N_k) - \log(\sum_{j=1}^{n_s} N_j)) \ , \tag{16}$$

où $g_k$ est une constante à $p$ et $T$ fixées appelée enthalpie libre molaire de l'espèce $k$.

Le problème mathématique à résoudre est donc de minimiser la fonction $G$ donnée par (16) sous la contrainte (15).

Il s'agit d'abord de considérer *l'existence et l'unicité de la solution*. Pour cela, on se ramène à un théorème classique puisque les $N_k$ satisfont :

$$0 \leq N_k \leq N_k^{sup} := \min_j(N_j^e/E_{kj}) \ .$$

(pourquoi ?) Ceci (avec (15)) définit un compact de $\mathbb{R}^{n_s}$ et on vérifie facilement que $G$ est continue sur ce compact.

Pour l'unicité, on procède en deux temps : on calcule le gradient de $G$ dont les composantes valent :

$$\frac{\partial G}{\partial N_k} = g_k + \log(X_k) \ .$$

On vérifie d'abord que le minimum n'est pas atteint sur un point de la frontière du domaine de contraintes (c'est-à-dire pour $N_k = 0$ ou $N_k = N_k^{sup} := \min_j(N_j^e/E_{kj})$. Pour cela, on montre que la dérivée normale tend vers $-\infty$ en un point où $N_k = 0$ et un tel point existe dans les deux cas (même le second).

Puisque le point de minimum est atteint à l'intérieur, il s'agit de voir que $G$ est strictement convexe *sur la contrainte*. Or on a :

$$\frac{\partial^2 G}{\partial N_k \partial N_j} = \delta_{k,j}(N_k)^{-1} - (\bar{N})^{-1} \ ,$$

et il faut tester $D^2G$ suivant les directions possibles $\Delta N$ le long de la contrainte, pour lesquelles on a $\Delta N.E = 0$ et donc $\Delta N$ change de signe. On a :

$$D^2G(N)\Delta N \cdot \Delta N = \sum_{k=1}^{n_s} \frac{[(\Delta N)_k]^2}{N_k} - \left(\sum_{k=1}^{n_s}(\Delta N)_k\right)^2 /\bar{N} \ .$$

Par Cauchy-Schwarz, cette quantité est strictement positive puisque $\Delta N$ ne peut être colinéaire au vecteur $N$.

*Pour calculer numériquement* l'équilibre, il est plus simple d'utiliser que la contrainte est affine et d'introduire une base $(f_j)_j$ de l'espace vectoriel associé (donc une base du noyau de la transposée de $E$). En écrivant :

$$N = N_0 + x_1.f_1 + ... + x_l.f_l \ ,$$

(que vaut $l$ ?), on se ramène à un problème d'optimisation sans contrainte sur $(x_1, \cdots, x_l)$ pour lequel on peut appliquer des méthodes classiques (gradient conjugué, Newton,...).

**Exercice 10.** *Traiter complètement le cas du mélange $H_2$-$O_2$-$H_2O$ (en prenant des $g_k$ égaux respectivement à $-18$, $-27$, $-45$).*
*Puis celui faisant aussi intervenir les espèces $OH$, $O$ et $H$ (en prenant pour ces espèces des $g_k$ égaux respectivement à $-21$, $-2$, $+2$).*

# 5 Appendice : Compléments

## 5.1 Étude générale des méthodes à un pas

On conserve, dans cette section, une grille uniforme de pas $h = \frac{T}{N}$. Les méthodes à un pas sont des méthodes de la forme :

$$\left\{ \begin{array}{rcl} y_{i+1} & = & y_i + h\Phi(t_i, y_i, h) \\ y_0 & = & y_{0,h} \end{array} \right.$$

où $\Phi$ est une fonction continue sur $[0,T] \times \mathbb{R}^n \times [0,H]$, $H$ désignant un pas de discrétisation maximal.

### 5.1.1 Propriétés importantes d'une méthode à un pas

- **CONSISTANCE**

**Définition 1.** *On appelle erreur de consistance de la méthode à un pas, la quantité :*

$$\Sigma_h = \sum_{i=0}^{N-1} |y(t_{i+1}) - y(t_i) - h\Phi(t_i, y(t_i), h)| \ .$$

*La méthode est dite consistante si $\Sigma_h \to 0$ quand $h \to 0$.*

La quantité $y(t_{i+1}) - y(t_i) - h\Phi(t_i, y(t_i), h)$ est l'analogue de ce que nous avons noté $\varepsilon_i$ ci-dessus ; nous conserverons au besoin cette notation.

- **STABILITÉ**

**Définition 2.** *La méthode à un pas est dite stable s'il existe deux constantes $S_1, S_2$ telles que, si $(\tilde{y}_i)_i$ est défini par :*

$$\left\{ \begin{array}{rcl} \tilde{y}_{i+1} & = & \tilde{y}_i + h\Phi(t_i, \tilde{y}_i, h) + \varepsilon_i \\ \tilde{y}_0 & = & \tilde{y}_{0,h} \end{array} \right. ,$$

*alors :*

$$\max_i |y_i - \tilde{y}_i| \leq S_1 |y_{0,h} - \tilde{y}_{0,h}| + S_2 \sum_{i=0}^{N-1} |\varepsilon_i| \ .$$

Bien entendu, la méthode est dite **convergente** si $\max_i |y_i - y(t_i)| \to 0$ quand $h \to 0$.

**Théorème 6.** *Toute méthode stable et consistante converge à condition que $y_{0,h} \to y(0)$ quand $h \to 0$.*

La dernière condition étant toujours satisfaite en pratique, l'étude de la convergence des méthodes à un pas se réduit à l'étude de leur consistance et de leur stabilité, ce qui est plus simple comme on va le voir.
**Preuve :** Si on note $\tilde{y}_i = y(t_i)$, on a par définition de $\varepsilon_i$ (juste après la définition de la consistance) :

$$\tilde{y}_{i+1} = \tilde{y}_i + h\Phi(t_i, \tilde{y}_i, h) + \varepsilon_i \ .$$

Et $\tilde{y}_0 = y(0)$. Puisque la méthode est stable :

$$\max_i |y_i - y(t_i)| \leq S_1 |y_{0,h} - y(0)| + S_2 \sum_{i=0}^{N-1} |\varepsilon_i| \ .$$

Hors, les deux quantités du membre de droite tendent vers 0 quand $h$ tend vers 0 par consistance, donc le résultat est acquis. □

### 5.1.2 Condition nécessaire et suffisante de consistance

**Théorème 7.** *La méthode à un pas est consistante si et seulement si :*

$$\Phi(t, z, 0) = f(t, z) \ ,$$

*pour tous* $t \in [0, T]$, $z \in \mathbb{R}^n$.

**Preuve :** On ne va vraiment détailler que la condition suffisante.

$$\begin{aligned} \varepsilon_i &= y(t_{i+1}) - y(t_i) - h\Phi(t_i, y(t_i), h) \\ &= \int_{t_i}^{t_{i+1}} [f(s, y(s)) - \Phi(t_i, y(t_i), h)] ds \ . \end{aligned}$$

Mais, pour $s \in [t_i, t_{i+1}]$ :

$$|f(s, y(s)) - \Phi(t_i, y(t_i), h)| \leq |f(s, y(s)) - f(t_i, y(t_i))| +$$

$$|f(t_i, y(t_i)) - \Phi(t_i, y(t_i), 0)| + |\Phi(t_i, y(t_i), 0) - \Phi(t_i, y(t_i), h)|$$

Le premier et le troisième terme sont des termes petits (uniformément en $i$) par l'uniforme continuité de $f$, $y$ et $\Phi$ ; on les estime par un $\delta(h)$ qui tend vers 0 avec $h$.

Si le terme du milieu est nul (condition de consistance) alors $|\varepsilon_i| \leq h\delta(h)$ et $\Sigma_h \leq Nh\delta(h) = T\delta(h) \to 0$ quand $h \to 0$. D'où la consistance.
NB : la condition suffisante se prouve en examinant la preuve d'un peu plus près : si le terme du milieu n'est pas nul... □

### 5.1.3 Condition suffisante de stabilité

**Théorème 8.** *La méthode à un pas est stable si $\Phi(t, z, h)$ est lipschitzien en $z$ pour tous $t \in [0, T]$, $h \in [0, H]$ avec une constante de lipschitz indépendante de $t$ et $h$.*

**Preuve :** On note $\theta_i = |y_i - \tilde{y}_i|$. On a :

$$\theta_{i+1} = |y_i + h\Phi(t_i, y_i, h) - \tilde{y}_i - h\Phi(t_i, \tilde{y}_i, h) - \varepsilon_i| \ .$$

D'où :

$$\theta_{i+1} \leq \theta_i + h|\Phi(t_i, y_i, h) - \Phi(t_i, \tilde{y}_i, h)| + |\varepsilon_i| \ .$$

Si $\Phi$ est lipschitzienne en sa deuxième variable de constante de lipschitz $\tilde{L}$ :

$$\theta_{i+1} \leq (1 + \tilde{L}h)\theta_i + |\varepsilon_i| \ .$$

Une petite amélioration du Lemme de Gronwall discret nous donne :

$$\theta_{i+1} \leq \exp(\tilde{L}(i+1)h)\theta_0 + \sum_{k=0}^{i} \exp(\tilde{L}(i-k)h)|\varepsilon_k| \ .$$

En majorant $\exp(\tilde{L}(i+1)h)$ et $\exp(\tilde{L}(i-k)h)$ par $\exp(\tilde{L}T)$, on a le résultat avec $S_1 = S_2 = \exp(\tilde{L}T)$. □

### 5.1.4 Ordre d'un schéma

La question que l'on se pose ici est la suivante : peut-on avoir une meilleure précision que dans le cas de la méthode d'Euler en choisissant bien la méthode à un pas et comment faut-il la choisir ?

**Définition 3.** *On dit qu'une méthode à un pas est d'ordre $p \geq 1$ si, pour toute solution de l'EDO, il existe une constante $C > 0$ telle que :*

$$|\varepsilon_i| = |y(t_{i+1}) - y(t_i) - h\Phi(t_i, y(t_i), h)| \leq Ch^{p+1} \ .$$

Pourquoi le "$p + 1$" dans l'ordre $p$ ? Deux raisons concourantes :

1. $|\dfrac{y(t_{i+1}) - y(t_i)}{h} - \Phi(t_i, y(t_i), h)| \leq Ch^p$ et la quantité considérée approche l'équation à l'ordre $p$.
2. $\Sigma_h = \displaystyle\sum_{i=0}^{N-1} |y(t_{i+1}) - y(t_i) - h\Phi(t_i, y(t_i), h)| \leq Ch^p$, donc ordre $p$ = erreur en $h^p$.

Cette dernière idée est justifiée par le résultat suivant dont la preuve est immédiate à partir du résultat de convergence :

**Corollaire 1.** *Si la méthode à un pas est d'ordre $p \geq 1$ et si $|y(0) - y_{0,h}| \leq \tilde{C}h^p$ alors $\max_i |y_i - y(t_i)| \leq \hat{C}h^p$.*

Nous donnons maintenant une condition nécessaire et suffisante pour qu'une méthode à un pas soit d'ordre $p$ dans $\mathbb{R}$. Pour cela, nous commençons par un résultat de régularité sur la solution $y$ de l'EDO : c'est une règle générale, on ne sait bien approcher que des fonctions régulières et donc nous aurons besoin d'une certaine régularité de la solution. Nous formulons ce résultat en dimension 1, c'est-à-dire quand $y$ est à valeurs dans $\mathbb{R}$.

**Théorème 9.** *Si la fonction $f$ est de classe $C^p$ sur $[0,T] \times \mathbb{R}$ alors $y$ est de classe $C^{p+1}$ et :*

$$y^{(k+1)}(t) = f^{[k]}(t, y(t)) \quad \textit{dans } ]0,T[ \qquad (k = 0, 1, \cdots, p),$$

*où les fonctions $f^{[k]}$ sont définies sur $[0,T]$ par :*

$$f^{[0]}(t,z) = f(t,z)\ ,$$

$$f^{[k+1]}(t,z) = \frac{\partial}{\partial t} f^{[k]}(t,z) + f(t,z)\frac{\partial}{\partial y} f^{[k]}(t,z)\ ,$$

*pour $k = 0, 1, \cdots, p-1$.*

La preuve se fait facilement et elle est laissée en exercice (où l'on pourra aussi réfléchir à la formulation de ce résultat si $y$ est à valeurs dans $\mathbb{R}^n$).

La condition nécessaire et suffisante pour qu'une méthode à un pas soit d'ordre $p$ dans $\mathbb{R}$ est donnée par le :

**Théorème 10.** *On suppose que $f$ est de classe $C^p$ et que, pour $k \leq p$, les dérivées partielles $\dfrac{\partial^k \Phi}{\partial h^k}$ existent et sont continues sur $[0,T] \times \mathbb{R} \times [0,H]$. Alors la méthode à un pas est d'ordre $p$ si, pour tous $t \in [0,T]$, $z \in \mathbb{R}$ :*

$$\begin{aligned}
\Phi(t,z,0) &= f(t,z) \\
\frac{\partial \Phi}{\partial h}(t,z,0) &= \frac{1}{2} f^{[1]}(t,z) \\
&\vdots \quad \vdots \quad \vdots \\
\frac{\partial^k \Phi}{\partial h^k}(t,z,0) &= \frac{1}{k+1} f^{[k]}(t,z), \quad k \leq p-1\ .
\end{aligned}$$

*Dans ce cas :*

$$|\varepsilon_i| = \frac{1}{i!} h^{p+1} \left( \frac{1}{p+1} f^{[p]}(t_i, y(t_i)) - \frac{\partial^p \Phi}{\partial h^p}(t_i, y(t_i), 0) \right) + o(h^{p+1})\ .$$

**Preuve :** Comme $f$ est de classe $C^p$, $y$ est de classe $C^{p+1}$ et $y^{(k)}(t) = f^{[k-1]}(t, y(t))$ par le résultat sur la régularité des solutions.

Par la formule de Taylor, on a :

$$\begin{aligned} y(t_{i+1}) - y(t_i) &= \sum_{k=1}^{p} \frac{h^k}{k!} y^{(k)}(t_i) + O(h^{p+1}) \\ &= \sum_{k=1}^{p} \frac{h^k}{k!} f^{[k-1]}(t_i, y(t_i)) + O(h^{p+1}) \end{aligned}$$

et :

$$\Phi(t_i, y(t_i), h) = \sum_{k=0}^{p} \frac{h^k}{k!} \frac{\partial^k \Phi}{\partial h^k}(t_i, y(t_i), 0) + o(h^p) \ .$$

Il en résulte que :

$$\begin{aligned} h\Phi(t_i, y(t_i), h) &= \sum_{k=0}^{p} \frac{h^{k+1}}{k!} \frac{\partial^k \Phi}{\partial h^k}(t_i, y(t_i), 0) + o(h^{p+1}) \\ &= \sum_{k=1}^{p+1} \frac{h^k}{(k-1)!} \frac{\partial^{k-1} \Phi}{\partial h^{k-1}}(t_i, y(t_i), 0) + o(h^{p+1}) \end{aligned}$$

et :

$$\varepsilon_i = \sum_{k=1}^{p} \left[ \frac{1}{k} f^{[k-1]}(t_i, y(t_i)) - \frac{\partial^{k-1} \Phi}{\partial h^{k-1}}(t_i, y(t_i), 0) \right] \frac{h^k}{(k-1)!} + O(h^{p+1}) \ .$$

Sous les hypothèses du théorème, tous les termes entre crochets sont nuls et le résultat est acquis.

On peut le préciser avec la formule de Taylor avec reste intégral, ce qui donne la deuxième partie du théorème. □

### 5.1.5 Exemples

On va chercher la meilleure fonction $\Phi$, c'est-à-dire celle qui donne le meilleur ordre de convergence, parmi celles qui sont de la forme :

$$\Phi(t, z, h) = a_1 f(t, z) + a_2 f(t + p_1 h, z + p_2 h f(t, z)) \ ,$$

où $a_1, a_2, p_1, p_2$ sont des paramètres à fixer "au mieux".

On a une méthode consistante d'ordre 1 si :

$$\Phi(t, z, 0) = f(t, z) \ ,$$

donc si :

$$a_1 f(t, z) + a_2 f(t, z) = f(t, z) \ ,$$

d'où la première condition $a_1 + a_2 = 1$.

Pour avoir de l'ordre 2, il faut que :

$$\frac{\partial \Phi}{\partial h}(t,z,0) = \frac{1}{2} f^{[1]}(t,z)) = \frac{1}{2}\left(\frac{\partial f}{\partial t}(t,z) + f(t,z)\frac{\partial f}{\partial y}(t,z)\right) .$$

Or :

$$\frac{\partial \Phi}{\partial h}(t,z,h) = a_2 p_1 \frac{\partial f}{\partial t}(t+p_1 h, z+p_2 h f(t,z)) + a_2 p_2 f(t,z)\frac{\partial f}{\partial y}(t+p_1 h, z+p_2 h f(t,z)) .$$

La deuxième condition est donc $a_2 p_1 = a_2 p_2 = \frac{1}{2}$.

Par contre, on vérifie facilement (le faire !) que l'on ne peut pas aller plus loin. En prenant, $a_2 = \alpha$ comme paramètre, on a une famille de méthodes à un pas d'ordre 2 :

$$\Phi(t,z,h) = (1-\alpha) f(t,z) + \alpha f(t + \frac{h}{2\alpha}, z + \frac{h}{2\alpha} f(t,z)) .$$

Ces méthodes sont connues sous le nom de :

— $\alpha = 1$, méthode de la tangente améliorée,
— $\alpha = 1/2$, méthode d'Euler modifiée,
— $\alpha = 1$, méthode de Heun.