COURS DE DEA Niveau II

Intégration symplectique des systèmes hamiltoniens intégrables : comportement en temps long

Philippe Chartier et Erwan Faou

10-12 Juin 2003

Ces notes de cours ont été élaborées dans le but de comprendre les détails de l'application de la théorie KAM (Kolmogorov, Arnold, Moser) aux schémas symplectiques appliqués aux systèmes dynamiques Hamiltoniens. Le contenu mathématique repose essentiellement sur le livre *Geometric Numerical Integration, Structure-Preserving Algorithms for Ordinary Differential Equations* écrit par E. Hairer, C. Lubich et G. Wanner publié chez Springer et référencé ci-dessous. A partir des chapitres IX et X de cet ouvrage, le présent travail a simplement consisté à écrire à fond les démonstrations dans le but de les expliquer à des non-spécialistes du domaine.

Nous remercions vivement Ernst Hairer, Christian Lubich et Gerhard Wanner d'avoir accepté la diffusion de ces notes.

Références :

- V. I. Arnold, *Mathematical Methods of Classical Mechanics*, Second Edition, Springer-Verlag, 1989.
- E. Hairer, S.P. Norsett and G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff Problems*, Second Revised Edition, Springer-Verlag, 1993.
- J.M. Sanz-Serna and M.P. Calvo, *Numerical Hamiltonian Problems*, Chapman & Hall, 1994.
- E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems*, Second Revised Edition, Springer-Verlag, 1996.
- E. Hairer, *Numerical Geometric Integration*, support de cours, `http ://www.unige.ch/math/folks/hairer/polycop.html`, 1999.
- E. Hairer, Chr. Lubich and G. Wanner, *Geometric Numerical Integration, Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer-Verlag, 2002.
- E. Hairer and Chr. Lubich, *The life-span of backward error analysis for numerical integrators*, Numer. Math. 76 (1997) pp. 441-462.

Table des matières

Chapitre 1

Expériences numériques et objectifs du cours

1.1 Premiers exemples et expériences numériques

Dans ce chapitre, on présente un certain nombre d'expériences numériques, destinées à illustrer les aspects les plus remarquables des méthodes dites symplectiques, que l'on définira plus avant, lorsqu'elles sont appliquées aux systèmes différentiels Hamiltoniens, c'est-à-dire aux systèmes de la forme :

$$\begin{array}{rcl} \dot{p} &=& -\nabla_q H \\ \dot{q} &=& \nabla_p H \end{array}$$

où H est une fonction de $\mathbb{R}^{2d}$ dans $\mathbb{R}$, dite Hamiltonienne. Notons tout de suite qu'il n'est pas rare que H s'écrive $H(p,q) = \frac{1}{2}p^T M^{-1} p + V(q)$, c'est-à-dire comme somme d'une énergie cinétique et d'une énergie potentielle. C'est le cas des exemples présentés ci-dessous.

1.1.1 Exemples de systèmes Hamiltoniens

Oscillateur harmonique

L'hamiltonien est de la forme :

$$H(p,q) = \frac{1}{2}(p^2 + k^2 q^2)$$

Le système correspondant s'obtient par les équations :

$$\left\{ \begin{array}{rcccl} \dot{p} &=& -\frac{\partial H}{\partial q} &=& -k^2 q \\ \dot{q} &=& \frac{\partial H}{\partial p} &=& p \end{array} \right. .$$

La trajectoire exacte dans l'espace des phases (p,q) est une ellipse dépendant des conditions initiales.

Problème de Kepler dans le plan

$$\begin{array}{rcl} H(p^1,p^2,q^1,q^2) &=& \frac{1}{2}[(p^1)^2 + (p^2)^2] - \frac{1}{\sqrt{(q^1)^2 + (q^2)^2}}, \\ &=& T(p) + V(q). \end{array}$$

Le système correspondant peut s'écrire comme un système d'ordre 2 :

$$\left\{\begin{array}{lllll} \dot{p} & = & -\frac{\partial H}{\partial q} & = & -V'(q) \\ \dot{q} & = & \frac{\partial H}{\partial p} & = & p \end{array}\right. \Longleftrightarrow \ddot{q} = -V'(q)$$

La trajectoire dans l'espace des phases (p_1, p_2) est une ellipse dans le cas étudié.

Évolution d'un cristal d'Argon (dans le plan)

$$\begin{aligned} H(p,q) &= \frac{1}{2}\sum_{i=1}^{N} m_i^{-1} p_i^T p_i + \sum_{i=2}^{N}\sum_{j=1}^{i-1} V_{ij}(\|q_i - q_j\|), \\ &= T(p) + V(q), \end{aligned}$$

où les V_{ij} sont des fonctions potentiel, les q_i désignent les positions des atomes, et les q_i les quantités de mouvement des atomes. Pour ce test on prend le potentiel de Lennard-Jones

$$V_{ij} = 4\varepsilon_{ij}\left((\frac{\sigma_{ij}}{r})^{12} - (\frac{\sigma_{ij}}{r})^{6} \right),$$

et $\varepsilon_{ij} = \varepsilon = 119.8 k_B [J]$, $m_i = m = 66.34\, 10^{-27} [kg]$, $\sigma_{ij} = 0.341 [nm]$ et $N = 7$. La température du système est définie par :

$$T = \frac{1}{2Nk_B}\sum_{i=1}^{N} m_i \|\dot{q}_i\|^2.$$

L'équation différentielle correspondante est donc de la forme :

$$\begin{array}{lllll} \dot{p} & = & -\frac{\partial H}{\partial q} & = & -V'(q), \\ \dot{q} & = & \frac{\partial H}{\partial p} & = & M^{-1}p. \end{array}$$

1.1.2 Schémas numériques élémentaires pour l'oscillateur harmonique

Méthode d'Euler explicite

$$\left\{\begin{array}{lllllll} p_{n+1} & = & p_n + h(-k^2 q_n) & = & p_n & + & (-hk^2) q_n \\ q_{n+1} & = & q_n + h(p_n) & = & hp_n & + & q_n \end{array}\right.$$

Méthode d'Euler implicite (ou Euler rétrograde)

$$\left\{\begin{array}{lllllll} p_{n+1} & = & p_n + h(-k^2 q_{n+1}) & = & \frac{1}{1+h^2k^2} p_n & + & \frac{-hk^2}{1+h^2k^2} q_n \\ q_{n+1} & = & q_n + h(p_{n+1}) & = & \frac{h}{1+h^2k^2} p_n & + & \frac{1}{1+h^2k^2} q_n \end{array}\right.$$

Méthode de Runge d'ordre 2

$$\left\{\begin{array}{lll} p_{n+1/2} & = & p_n + \frac{h}{2}(-k^2 q_n) \\ q_{n+1/2} & = & q_n + \frac{h}{2}(p_n) \\ p_{n+1} & = & p_n + h(-k^2 q_{n+1/2}) \\ q_{n+1} & = & q_n + h(p_{n+1/2}), \end{array}\right.$$

Méthode du point milieu (ou Gauss d'ordre 2)

$$\left\{\begin{array}{lclcl} p_{n+1} & = & p_n + h(-k^2 \frac{q_{n+1}+q_n}{2}) & = & \frac{1-h^2k^2/4}{1+h^2k^2/4} p_n \quad + \quad \frac{-hk^2}{1+h^2k^2/4} q_n \\ q_{n+1} & = & q_n + h(\frac{p_{n+1}+p_n}{2}) & = & \frac{h}{1+h^2k^2/4} p_n \quad + \quad \frac{1-h^2k^2/4}{1+h^2k^2/4} q_n \end{array}\right.$$

Méthode (partitionnée) d'Euler symplectique

$$\left\{\begin{array}{lcl} p_{n+1} & = & p_n - hV'(q_n), \\ q_{n+1} & = & q_n + hp_{n+1}, \end{array}\right. .$$

Méthode de Störmer-Verlet

$$\begin{array}{rcl} p_{n+1/2} & = & p_n - \frac{h}{2} V'(q_n), \\ q_{n+1} & = & q_n + hM^{-1} p_{n+1/2}, \\ p_{n+1} & = & p_{n+1/2} - \frac{h}{2} V'(q_{n+1}) \end{array}$$

1.1.3 Expériences numériques

Comportement qualitatif pour l'oscillateur harmonique

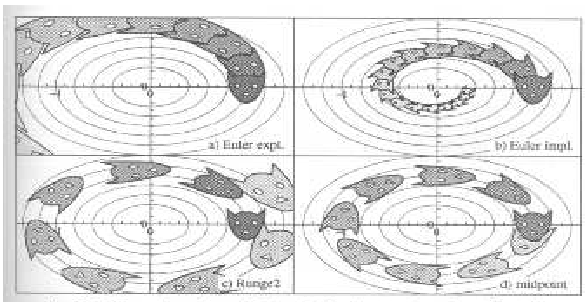


FIG. 1.1 – Expériences numériques d'après Hairer-Wanner volume 1

Comparaison des trajectoires du système de Kepler

Conservation de l'énergie - Erreur Globale pour le problème de Kepler

Conservation de l'énergie et de la température pour le problème de cristal d'Argon

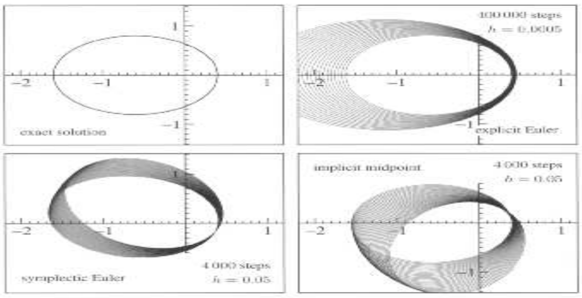


FIG. 1.2 – D'après Hairer, Geometric Integration, pp. 5

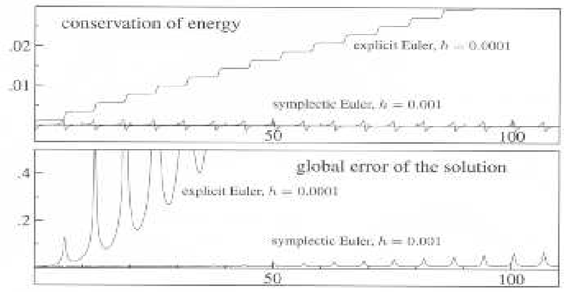


FIG. 1.3 – D'après Hairer, Geometric Integration, pp. 6

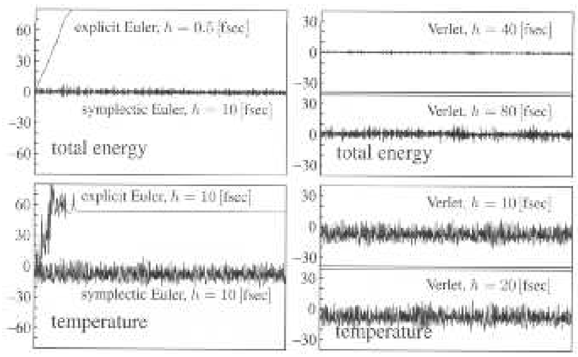


FIG. 1.4 – D'après Hairer, Geometric Integration, pp. 11

1.2 Objectifs du cours

Les expériences numériques présentées ci-avant mettent en évidence les particularités suivantes :
- conservation de l'Hamiltonien
- croissance linéaire de l'erreur globale
- comportement qualitatif "conforme à la physique sous-jacente"

L'objectif de ce cours est d'apporter des éléments d'explication au comportement favorable des schémas numériques symplectiques en temps long. Après un bref passage en revue des méthodes de Runge-Kutta symplectiques, illustré par diverses expériences numériques, nous nous attache-

1638	Galilée	Calcul d'orbites
1687	Newton	$m\ddot{x} = F(x)$
1760-1788	Lagrange	$L = T - U$ et $\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) = \frac{\partial L}{\partial q}$
1834	Hamilton	$H = T + U$ et $\dot{p} = -\nabla_q H,\ \dot{q} = \nabla_p H$
1842	Jacobi	EDP du premier ordre pour les fonctions génératrices
1892-1899	Poincaré	Mécanique céleste (3 volumes monumentaux)
1960	Siegel	Condition diophantienne : $\lvert\alpha \cdot \omega\rvert \geq \gamma\lvert\alpha\rvert^{-\nu}$
1963	Arnold	Preuve du théorème KAM
1983	Ruth	Premier intégrateur symplectique
1988	Sanz-Serna	Caractérisation des méthodes de Runge-Kutta symplectiques
1994	Benettin	Analyse rétrograde

rons à présenter les deux ingrédients principaux de l'intégration géométrique : d'une part l'analyse rétrograde, d'autre part la théorie des perturbations de systèmes hamiltoniens intégrables (KAM). On montrera enfin comment la combinaison de ces deux outils permet d'établir les résultats de conservation d'energie et de préservation des tores invariants observés pour les méthodes décrites en introduction.

Le contenu de ce cours est très largement inspiré du livre de E. Hairer, C. Lubich et G. Wanner, intitulé "Geometric Numerical Integration" et disponible chez Springer.

Chapitre 2

Systèmes hamiltoniens : équations canoniques et propriétés

A partir des équations de Lagrange :

$$\begin{aligned} L &= T(q,\dot q) - U(q), && (2.0.1)\\ \frac{d}{dt}\left(\frac{\partial L}{\partial \dot q}\right) &= \frac{\partial L}{\partial q}, && (2.0.2)\end{aligned}$$

on pose (transformation de Legendre) :

$$\begin{aligned} p &= \nabla_{\dot q} L(q,\dot q), && (2.0.3)\\ H &= p^T \dot q - L(q,\dot q). && (2.0.4)\end{aligned}$$

Théorème 1 *Si la transformation $\dot q \to p = \nabla_{\dot q} L(q,\dot q)$ est un difféomorphisme pour tout q, alors les équations de Lagrange et les équations de Hamilton :*

$$\begin{aligned} \dot p &= -\nabla_q H(p,q), && (2.0.5)\\ \dot q &= \nabla_p H(p,q), && (2.0.6)\end{aligned}$$

sont équivalentes.

Preuve : En différentiant H on obtient :

$$\begin{aligned} \frac{\partial H}{\partial p} &= \dot q^T + p^T \frac{\partial \dot q}{\partial p} - \frac{\partial L}{\partial \dot q}\frac{\partial \dot q}{\partial p} = \dot q^T, && (2.0.7)\\ \frac{\partial H}{\partial q} &= p^T \frac{\partial \dot q}{\partial q} - \frac{\partial L}{\partial q} - \frac{\partial L}{\partial \dot q}\frac{\partial \dot q}{\partial q} = -\frac{\partial L}{\partial q}. && (2.0.8)\end{aligned}$$

Donc :

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot q}\right) = \frac{d}{dt} p^T = -\frac{\partial H}{\partial q}. \qquad (2.0.9)$$

□

Pour p et q dans $\mathbb{R}^d$, et H une fonction de $\mathbb{R}^{2d}$ à valeurs réelles, suffisamment régulière (au moins C^2), on définit le système différentiel suivant :

$$\left\{ \begin{array}{rcl} \dot p &=& -\nabla_q H \\ \dot q &=& \nabla_p H \end{array} \right. .$$

En notant

$$y = \begin{pmatrix} p \\ q \end{pmatrix}, \; J = \begin{pmatrix} 0 & I_d \\ -I_d & 0 \end{pmatrix}$$

on peut aussi écrire

$$\dot{y} = J^{-1}\nabla H.$$

Notons tout de suite que l'on a les égalités suivantes :

$$J^T = -J, J^2 = -I_{2d} \text{ et } J^{-1} = -J. \tag{2.0.10}$$

2.1 Conservation de l'hamiltonien

Le long de toute trajectoire exacte :

$$\begin{aligned} \frac{d}{dt}H(y) &= \frac{\partial H}{\partial y}\frac{dy}{dt}, \\ &= (\nabla H)^T J^{-1}\nabla H, \\ &= 0, \end{aligned}$$

car la matrice J est antisymétrique, i.e. $J^T = -J$.

2.2 Conservation du volume

On définit la fonction "flot" φ_t, de l'espace des phases $\mathbb{R}^n$ dans lui-même, comme étant la fonction qui associe à $y_0 \in \mathbb{R}^n$ la valeur au temps t de la solution exacte du système $\dot{y} = f(y)$ passant par y_0 au temps $t = 0$. Pour un système hamiltonien $\varphi_t(p_0, q_0) = (p(t), q(t))$, où $p(0) = p_0$ et $q(0) = q_0$. On a alors :

Théorème 2 *Pour un système différentiel de la forme $\dot{y} = f(y)$, avec f de classe C^1 telle que $divf \equiv 0$, alors*

$$Vol(\varphi_t(A)) = Vol(A)$$

pour tout ensemble $A \subset \mathbb{R}^n$.

Preuve : Soit $\Psi_t(y) = \frac{\partial \varphi_t}{\partial y}(y)$. Ψ_t est solution de l'équation variationnelle

$$\frac{d}{dt}\Psi_t(y) = \frac{\partial f}{\partial y}(\varphi_t(y))\Psi_t(y), \tag{2.2.1}$$

$$\Psi_0(y) = I_{\mathbb{R}^n}. \tag{2.2.2}$$

Pour une matrice $M \in GL_n(\mathbb{R})$, on a :

$$\forall H \in GL_n(\mathbb{R}), \; (d_M \det)H = (\det M)Tr(M^{-1}H) \tag{2.2.3}$$

En effet :

$$\frac{\det(M + tH) - \det(M)}{t} = \frac{t^n \det(M) \det(\frac{1}{t}I_{\mathbb{R}^n} + M^{-1}H) - \det(M)}{t} = \det(M)\, Tr(M^{-1}H) + \mathcal{O}(t)$$

Il vient donc[1] :

$$\begin{aligned}
\frac{d}{dt}\det(\Psi_t(y)) &= (d_\Psi \det)\frac{d}{dt}\Psi_t = \det(\Psi_t)\,Tr(\Psi_t^{-1}\frac{d}{dt}\Psi_t), && (2.2.5)\\
&= \det(\Psi_t)\,Tr(\Psi_t^{-1}\,(\partial_y f(\varphi_t(y)))\,\Psi_t), && (2.2.6)\\
&= \det(\Psi_t)\,Tr(\partial_y f(\varphi_t(y))) = 0. && (2.2.7)
\end{aligned}$$

Ainsi, $\det(\Psi_t(y)) = \det(\Psi_0(y)) = \det(I_{\mathbb{R}^n}) = 1$, et

$$\int_{\varphi_t(A)} dz = \int_A \det((\partial_y\varphi_t)(y))dy = \int_A dy. \quad (2.2.8)$$

□

Pour un système Hamiltonien :

$$\mathrm{div}f = Tr(\frac{\partial f}{\partial y}) = Tr(J^{-1}\nabla^2 H) = Tr(\nabla^2 H J^{-T}) = -Tr(J^{-1}\nabla^2 H) = -\mathrm{div}f = 0.$$

2.3 Quelques éléments de géométrie

On considère le parallélogramme P de $\mathbb{R}^{2d}$ engendré par les vecteurs

$$\xi = \begin{bmatrix} \xi^p \\ \xi^q \end{bmatrix} \text{ et } \eta = \begin{bmatrix} \eta^p \\ \eta^q \end{bmatrix}$$

dans l'espace des "phases" (p, q) :

$$P = \{t\xi + s\eta \,|\, 0 \le t, s \le 1\}.$$

En dimension 1, l'aire *orientée* de P s'écrit :

$$\text{aire.orientée}(P) = \begin{vmatrix} \xi^p & \eta^p \\ \xi^q & \eta^q \end{vmatrix} = \xi^p\eta^q - \xi^q\eta^p.$$

En dimension $d > 1$, on remplace cette expression par la somme $\omega(\xi, \eta)$ des aires orientées des *projections* sur les plans (p_i, q_i) de P :

$$\omega(\xi, \eta) = \sum_{i=1}^{d} \begin{vmatrix} \xi_i^p & \eta_i^p \\ \xi_i^q & \eta_i^q \end{vmatrix} = \sum_{i=1}^{d} (\xi_i^p\eta_i^q - \xi_i^q\eta_i^p).$$

ω définit ainsi une *forme bilinéaire antisymétrique*, que l'on peut encore écrire :

$$\omega(\xi, \eta) = \xi^T J\eta,$$

où J est la matrice définie précédemment.

[1] On peut aussi faire un calcul direct. Soient $\Psi_1, \ldots, \Psi_n$ les vecteurs colonnes de Ψ et $\alpha_1, \ldots, \alpha_n$ ceux de $\alpha = \Psi^{-1}\partial_y f(\varphi_t(y))\Psi$. On a bien sûr pour tout $j = 1, \ldots, n$

$$\frac{d}{dt}\Psi_j = (\partial_y f(\varphi_t(y)))\Psi_j = \sum_{i=1}^{n} \alpha_{i,j}\Psi_i. \quad (2.2.4)$$

Le déterminant étant une n-forme antisymétrique ω^n, on a :

$$\begin{aligned}
\frac{d}{dt}\det(\Psi) &= \sum_{j=1}^{n} \omega^n(\Psi_1, \ldots, \Psi_{j-1}, \dot{\Psi}_j, \Psi_{j+1}, \ldots, \Psi_n) = \sum_{i,j=1}^{n} \omega^n(\Psi_1, \ldots, \Psi_{j-1}, \alpha_{i,j}\Psi_i, \Psi_{j+1}, \ldots, \Psi_n),\\
&= \sum_{j=1}^{n} \alpha_{j,j}\,\omega^n(\Psi_1, \ldots, \Psi_{j-1}, \Psi_j, \Psi_{j+1}, \ldots, \Psi_n) = Tr(\alpha)\det(\Psi) = \det(\Psi)Tr(\partial_y f(\varphi_t(y))) = 0.
\end{aligned}$$

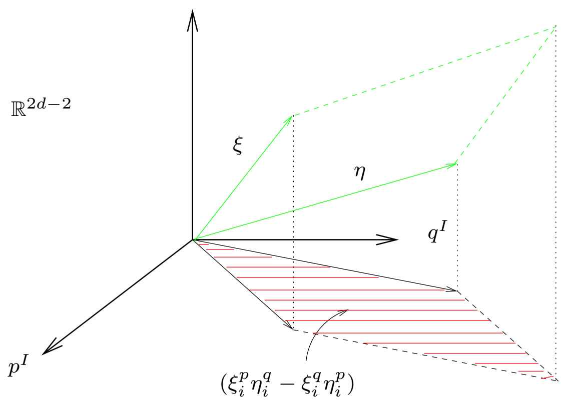


FIG. 2.1 – Application ω

2.4 Transformations symplectiques

Définition 1 *Une application linéaire* $A\ : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ *est dite symplectique si :*

$$A^T JA = J,$$

c'est-à-dire de manière équivalente, si :

$$\forall(\xi,\eta) \in \mathbb{R}^{2d} \times \mathbb{R}^{2d},\ \omega(A\xi, A\eta) = \omega(\xi,\eta).$$

L'ensemble des matrices satisfaisant $A^T JA = J$ est un sous-groupe de $GL_n(\mathbb{R})$ dit *groupe symplectique*. En dimension $d = 1$, la symplecticité de A ne traduit rien d'autre que la conservation des aires. En dimension $d > 1$, elle traduit la conservation de la somme des aires orientées des projections sur les plans (p_i, q_i).

Définition 2 *Une application* g *de* U *ouvert de* $\mathbb{R}^{2d}$ *dans* $\mathbb{R}^{2d}$*, de classe* C^1 *sur* U*, est dite symplectique si sa matrice jacobienne* $g'(p,q)$ *est symplectique pour tout* (p,q) *de* U*, c'est-à-dire si :*

$$\forall(p,q) \in U,\ \left(g'(p,q)\right)^T Jg'(p,q) = J,$$

ou de manière équivalente si :

$$\forall(p,q) \in U, \forall(\xi,\eta) \in \mathbb{R}^{2d} \times \mathbb{R}^{2d},\ \omega(g'(p,q)\xi, g'(p,q)\eta) = \omega(\xi,\eta).$$

Soit M une sous-variété bidimensionnelle de U, telle qu'il existe une "carte globale" :

$$M = \psi(K),$$

où K est un compact de $\mathbb{R}^2$ et $\psi(s,t)$ un difféomorphisme de K dans M. M peut être vue comme la limite de l'union de "petits parallélogrammes" engendrés par les vecteurs

$$\frac{\partial\psi}{\partial s}ds \text{ et } \frac{\partial\psi}{\partial t}dt.$$

Alors, la somme des aires orientées des projections sur les plans (p_i, q_i) de tous ces parallélogrammes s'écrit :

$$\Omega(M) = \iint_K \omega(\frac{\partial\psi}{\partial s}(s,t), \frac{\partial\psi}{\partial t}(s,t))dsdt.$$

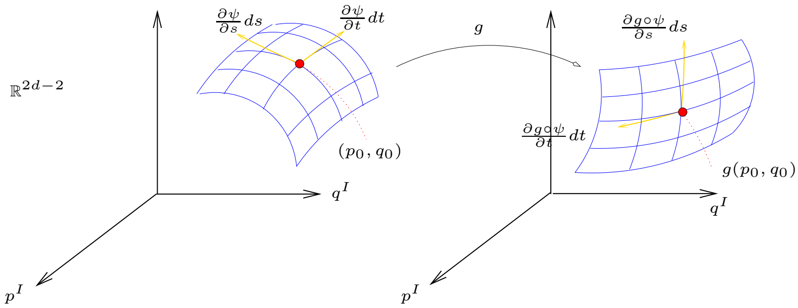


FIG. 2.2 – Image de M par g

Remarque 1 *La définition de $\Omega(M)$ ne dépend pas de la carte choisie. Soit par exemple $\varphi = (\varphi_1, \varphi_2)^T$ un difféomorphisme de $L \subset \mathbb{R}^2$ vers $K = \varphi(L) \subset \mathbb{R}^2$ de sorte que $M = (\psi \circ \varphi)(L)$. On a :*

$$\begin{aligned}
\iint_L \omega(\frac{\partial \psi\circ\varphi}{\partial u}(u,v), \frac{\partial \psi\circ\varphi}{\partial v}(u,v))dudv &= \iint_L \omega(\frac{\partial\varphi_1}{\partial u}\frac{\partial\psi}{\partial s} + \frac{\partial\varphi_2}{\partial u}\frac{\partial\psi}{\partial t}, \frac{\partial\varphi_1}{\partial v}\frac{\partial\psi}{\partial s} + \frac{\partial\varphi_2}{\partial v}\frac{\partial\psi}{\partial t})dudv, \\
&= \iint_L \frac{\partial\varphi_1}{\partial u}\frac{\partial\varphi_2}{\partial v}\omega(\frac{\partial\psi}{\partial s}, \frac{\partial\psi}{\partial t}) + \frac{\partial\varphi_2}{\partial u}\frac{\partial\varphi_1}{\partial v}\omega(\frac{\partial\psi}{\partial t}, \frac{\partial\psi}{\partial s})dudv, \\
&= \iint_L \det(\varphi'(u,v))\omega(\frac{\partial\psi}{\partial s}, \frac{\partial\psi}{\partial t})dudv, \\
&= \iint_K \omega(\frac{\partial\psi}{\partial s}, \frac{\partial\psi}{\partial t})dsdt.
\end{aligned}$$

Théorème 3 *Si $g\ : U \to \mathbb{R}^{2d}$, de classe C^1, est symplectique sur U, alors elle préserve $\Omega(M)$, c'est-à-dire :*

$$\Omega(g(M)) = \Omega(M).$$

Preuve : La sous-variété $g(M)$ peut être paramétrée par $g \circ \psi$ sur K. On a donc :

$$\begin{aligned}
\Omega(g(M)) &= \iint_K \omega(\frac{\partial g\circ\psi}{\partial s}(s,t), \frac{\partial g\circ\psi}{\partial t}(s,t))dsdt, \\
&= \iint_K \omega(g'(\psi(s,t))\frac{\partial\psi}{\partial s}(s,t), g'(\psi(s,t))\frac{\partial\psi}{\partial t}(s,t))dsdt, \\
&= \iint_K \left(\frac{\partial\psi}{\partial s}(s,t)\right)^T \underbrace{\left(g'(\psi(s,t))\right)^T J g'(\psi(s,t))}_{=J} \frac{\partial\psi}{\partial t}(s,t)dsdt, \\
&= \Omega(M).
\end{aligned}$$

□

2.5 Application aux systèmes hamiltoniens

On suppose donc $\dot{y} = J^{-1}\nabla H(y)$ et on note $\varphi_t(p_0, q_0)$ la fonction flot associée (dès lors qu'elle est bien définie).

Théorème 4 (Poincaré 1899) *Soit $H(p,q)$ une fonction de classe C^2 d'un ouvert U de $\mathbb{R}^{2d}$ dans $\mathbb{R}^{2d}$. Alors pour tout t tel que φ_t existe, φ_t est une transformation symplectique.*

Preuve : Pour tout t tel que φ_t existe, $\frac{\partial \varphi_t}{\partial y_0}$ est solution de l'équation variationnelle :

$$\begin{aligned}\dot{\Psi} &= J^{-1}\nabla^2 H(\varphi_t(y_0))\Psi,\\ \Psi(0) &= I_{2d}.\end{aligned}$$

Or $\nabla^2 H(\varphi_t(y_0))$ est symétrique, i.e. $(\left(\nabla^2 H(\varphi_t(y_0))\right)^T = \nabla^2 H(\varphi_t(y_0))$. D'où :

$$\begin{aligned}\frac{d}{dt}\left(\left(\frac{\partial \varphi_t}{\partial y_0}\right)^T J\frac{\partial \varphi_t}{\partial y_0}\right) &= \left(\frac{\partial \dot{\varphi_t}}{\partial y_0}\right)^T J\frac{\partial \varphi_t}{\partial y_0} + \left(\frac{\partial \varphi_t}{\partial y_0}\right)^T J\frac{\partial \dot{\varphi_t}}{\partial y_0},\\ &= \left(\frac{\partial \varphi_t}{\partial y_0}\right)^T (\nabla^2 H)^T \underbrace{J^{-T}J}_{=-J^{-1}J=-I}\frac{\partial \varphi_t}{\partial y_0}\\ &\quad + \left(\frac{\partial \varphi_t}{\partial y_0}\right)^T \underbrace{JJ^{-1}}_{=I}(\nabla^2 H)\frac{\partial \varphi_t}{\partial y_0} = 0.\end{aligned}$$

En $t=0$,

$$\left(\frac{\partial \varphi_t}{\partial y_0}\right)^T J\frac{\partial \varphi_t}{\partial y_0} = I^T J I = J,$$

ce qui permet de conclure. □

Théorème 5 *Soit U un ouvert de $\mathbb{R}^{2d}$ supposé étoilé, et f une fonction C^1 de U dans $\mathbb{R}^{2d}$. Alors si $\varphi_t(y)\ : U \to \mathbb{R}^{2d}$, le flot associé à l'équation $\dot{y} = f(y)$ pour t suffisamment petit, est symplectique, $\dot{y} = f(y)$ est un système hamiltonien sur U, c'est-à-dire qu'il existe une fonction H de classe C^2 sur U telle que*

$$\forall y \in U,\ f(y) = J^{-1}\nabla H(y).$$

Preuve : Pour tout t tel que φ_t existe, $\frac{\partial \varphi_t}{\partial y_0}$ est solution de l'équation variationnelle :

$$\begin{aligned}\dot{\Psi} &= f'(\varphi_t(y_0))\Psi,\\ \Psi(0) &= I_{2d}.\end{aligned}$$

En différentiant, il vient :

$$0 = \frac{d}{dt}\left(\left(\frac{\partial \varphi_t}{\partial y_0}\right)^T J\frac{\partial \varphi_t}{\partial y_0}\right) = \left(\frac{\partial \varphi_t}{\partial y_0}\right)^T \left((f'(\varphi_t(y_0)))^T J + Jf'(\varphi_t(y_0))\right)\frac{\partial \varphi_t}{\partial y_0}.$$

En écrivant l'égalité pour $t=0$ et en tenant compte de $J^T = -J$, on voit que $Jf'(y_0)$ doit être symétrique pour tout $y = y_0 \in U$. Donc $Jf(y) = \nabla_y H(y)$ d'après le lemme d'intégrabilité. □

Lemme 1 (Lemme d'intégrabilité) *Soit D un ouvert de $\mathbb{R}^n$ et $f : D \to \mathbb{R}^n$ une fonction de classe C^1, telle que $f'(y)$ soit symétrique pour tout $y \in D$. Alors, pour tout y_0 de D, il existe un voisinage $\mathcal{V}(y_0)$ et une fonction H définie sur $\mathcal{V}(y_0)$ telle que :*

$$\forall\, y \in \mathcal{V}(y_0),\; f(y) = \nabla_y H(y). \tag{2.5.1}$$

Preuve : Soit $B_0 \subset D$ une boule de centre y_0 contenue dans D. On définit H sur B_0 par :

$$H(y) = \int_0^1 (y-y_0)^T f(y_0 + t(y-y_0))dt.$$

Il vient alors :

$$\begin{aligned}
\frac{\partial H}{\partial y_j}(y) &= \int_0^1 f_j(y_0+t(y-y_0)) + t(y-y_0)^T \frac{\partial f}{\partial y_j}(y_0+t(y-y_0))dt,\\
&= \int_0^1 f_j(y_0+t(y-y_0)) + t(y-y_0)^T (\nabla_y f_j)(y_0+t(y-y_0))dt,\\
&= \int_0^1 \frac{d}{dt}\left(t f_j(y_0+t(y-y_0))\right)dt,\\
&= f_j(y).
\end{aligned}$$

□

Remarque 2 *Lorsque l'ouvert considéré est étoilé par rapport à l'un de ses points $y_0 \in U$, alors on peut définir H sur tout U.*

Remarque 3 *Dans le cas général d'un ouvert quelconque, le résultat du théorème 5 est faux. Considérons par exemple le système :*

$$\left\{\begin{array}{rcl} \dot p &=& \frac{p}{p^2+q^2}\\ \dot q &=& \frac{q}{p^2+q^2}\end{array}\right. \tag{2.5.2}$$

défini sur $U = \{(p,q) \in \mathbb{R}^2 \;; (p,q) \neq (0,0)\}$. Pour $(p_0,q_0) \in U$, le flot s'écrit :

$$\varphi_t(p_0,q_0) = \alpha(t,r_0)\begin{bmatrix} p_0\\ q_0\end{bmatrix}, \tag{2.5.3}$$

avec $\alpha(t,r_0) = \sqrt{1+2t/r_0^2}$ et $r_0 = \sqrt{p_0^2+q_0^2}$. Sa dérivée $\frac{\partial\varphi_t}{\partial(p_0,q_0)}$ est de la forme :

$$\frac{\partial \varphi_t}{\partial(p_0,q_0)} = \begin{bmatrix} (\partial_{p_0}\alpha)p_0+\alpha & (\partial_{q_0}\alpha)p_0\\ (\partial_{p_0}\alpha)q_0 & (\partial_{q_0}\alpha)q_0+\alpha\end{bmatrix} = \begin{bmatrix} (\partial_{r_0}\alpha)\frac{p_0^2}{r_0}+\alpha & (\partial_{r_0}\alpha)\frac{p_0q_0}{r_0}\\ (\partial_{r_0}\alpha)\frac{p_0q_0}{r_0} & (\partial_{r_0}\alpha)\frac{q_0^2}{r_0}+\alpha\end{bmatrix}. \tag{2.5.4}$$

C'est une matrice symplectique si $\alpha^2 + \alpha r_0(\partial_{r_0}\alpha) = 1$, ce qui se vérifie par un simple calcul. Localement, on peut écrire le système comme un système Hamiltonien. Par exemple, dans le demi-plan $p>0$, on peut prendre $H(p,q) = -\arctan\frac{q}{p}$ et vérifier que :

$$-\nabla_q H(p,q) = \frac{1}{p}\frac{1}{1+\frac{q^2}{p^2}} = \frac{p}{p^2+q^2} = \dot p, \tag{2.5.5}$$

$$\nabla_p H(p,q) = \frac{q}{p^2}\frac{1}{1+\frac{q^2}{p^2}} = \frac{q}{p^2+q^2} = \dot q. \tag{2.5.6}$$

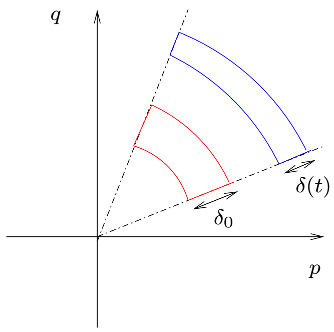


FIG. 2.3 – Symplecticité du flot de (2.5.2) : conservation de l'aire

A contrario, supposons qu'il existe un H de classe C^1 sur U, tel que le champ de vecteur $f(p,q) = (p^2+q^2)^{-1}[p,q]^T$ s'écrive $f(p,q) = J^{-1}\nabla_{p,q}H(p,q)$. Considérons alors la forme différentielle

$$\omega_{p,q} = f_2(p,q)dp - f_1(p,q)dq \tag{2.5.7}$$

et calculons son intégrale le long du chemin Γ paramétré par $(p,q) = (\cos(\theta), \sin(\theta))$:

$$\int_\Gamma \omega = \int_0^{2\pi} \omega_{\cos(\theta),\sin(\theta)} \begin{pmatrix} -\sin(\theta) \\ \cos(\theta) \end{pmatrix} d\theta = \int_0^{2\pi} (-\sin^2(\theta) - \cos^2(\theta))d\theta = -2\pi. \tag{2.5.8}$$

Or on aurait par ailleurs $\omega_{p,q} = (\partial_p H)(p,q)dp + (\partial_q H)(p,q)dq = dH$ dont l'intégrale sur Γ est nulle. H ne peut donc pas être défini sur tout U. L'hypothèse de simple connexité est essentielle pour cela.

L'une des propriétés fondamentales des transformations symplectiques est qu'elles préservent la structure Hamiltonienne. Plus encore, c'est une propriété caractéristique. On a en effet le théorème suivant :

Théorème 6 *Soit Ψ un difféomorphisme d'un ouvert U de $\mathbb{R}^{2d}$ dans $V = \Psi(U)$. Alors, si Ψ est symplectique, le système Hamiltonien*

$$\dot{y} = J^{-1}\nabla_y H(y)$$

s'écrit dans les variables $z = \Psi(y)$:

$$\dot{z} = J^{-1}\nabla_z K(z) \tag{2.5.9}$$

avec $K(z) = (H \circ \Psi^{-1})(z)$. Réciproquement, si Ψ transforme tout système d'Hamiltonien H en un système Hamiltonien d'Hamiltonien $K = H \circ \Psi^{-1}$, alors Ψ est symplectique.

Preuve : En dérivant z par rapport à t, il vient :

$$\dot{z} = \Psi'(y)\dot{y} = \Psi'(y)J^{-1}\nabla_y H(\Psi^{-1}(z)).$$

Or on a, en omettant les arguments des fonctions, $\nabla_y H = (\partial_y(K \circ \Psi))^T = ((\partial_z K)(\partial_y \Psi))^T = (\Psi')^T(\nabla_z K)$. D'où :

$$\dot{z} = \Psi'(\Psi^{-1}(z))J^{-1}(\Psi'(\Psi^{-1}(z)))^T \nabla_z K(z). \tag{2.5.10}$$

Si Ψ est symplectique, $(\Psi')^T J(\Psi') = J$, donc $\Psi' J^{-1}(\Psi')^T = J^{-1}$ et

$$\dot{z} = J^{-1}\nabla_z K(z).$$

Réciproquement, si le système en z (2.5.10) est hamiltonien pour tout K, alors $\Psi' J^{-1}(\Psi')^T = J^{-1}$ donc $(\Psi')^T J(\Psi') = J$ et Ψ est symplectique. □

Chapitre 3

Méthodes de Runge-Kutta symplectiques

Dans ce chapitre, on rappelle brièvement la forme générale des méthodes d'intégration du type Runge-Kutta, avant de caractériser celles d'entre elles qui sont symplectiques.

3.1 Méthodes numériques à un pas

Pour un système différentiel de la forme $\dot{y} = f(y)$, une méthode de Runge-Kutta s'écrit

$$\left\{ \begin{array}{rcl} Y_1 & = & y_n + h \sum_{j=1}^{s} a_{1j} \, f(Y_j) \\ & \vdots & \\ Y_s & = & y_n + h \sum_{j=1}^{s} a_{sj} \, f(Y_j) \\ y_{n+1} & = & y_n + h \sum_{j=1}^{s} b_j \, f(Y_j) \end{array} \right.$$

Elle est donc déterminée par la donnée des coefficients $A = (a_{ij}) \in \mathbb{R}^{s \times s}$ et $b = (b_j) \in \mathbb{R}^s$.

Pour un système différentiel de la forme

$$\left\{ \begin{array}{rcl} \dot{p} & = & f(p, q), \\ \dot{q} & = & g(p, q), \end{array} \right. \tag{3.1.1}$$

on peut utiliser deux jeux de coefficients :

$$\left\{ \begin{array}{rcl} P_i & = & p_n + h \sum_{j=1}^{s} a_{ij} \, f(P_j, Q_j), \;\; i = 1, \cdots, s \\ Q_i & = & q_n + h \sum_{j=1}^{s} \hat{a}_{ij} \, g(P_j, Q_j), \;\; i = 1, \cdots, s \\ p_{n+1} & = & p_n + h \sum_{j=1}^{s} b_j \, f(P_j, Q_j) \\ q_{n+1} & = & q_n + h \sum_{j=1}^{s} \hat{b}_j \, g(P_j, Q_j) \end{array} \right. \tag{3.1.2}$$

Définition 3 *Soit une méthode de Runge-Kutta partitionnée de coefficients A, $\hat{A}$, b, $\hat{b}$. Pour $|h|$ suffisamment petit, les équations précédentes définissent une fonction "flot numérique" Φ_h, telle que $y_{n+1} = \Phi_h(y_n)$ définit la solution approchée. Si cette fonction est symplectique dès lors que le système est hamiltonien, alors la méthode est dite symplectique.*

3.2 Caractérisation des méthodes symplectiques

Lemme 2 *Pour une méthode de Runge-Kutta (partitionnée ou non), le diagramme suivant commute :*

$$
\begin{array}{ccc}
\dot{y} = f(y),\, y(0) = y_0 & \overset{\partial/\partial_{y_0}}{\rightarrow} & \begin{array}{c}\dot{y} = f(y),\, y(0) = y_0 \\ \dot{\Psi} = f'(y)\Psi,\, \Psi(0) = I\end{array} \\
\downarrow & & \downarrow \\
\textit{méthode} & & \textit{méthode} \\
\downarrow & & \downarrow \\
\{y_n\} & \overset{\partial/\partial_{y_0}}{\rightarrow} & \{y_n, \Psi_n\}
\end{array}
$$

Preuve : Considérons une méthode de Runge-Kutta $\mathcal{R}$ non-partitionnée de coefficients A et b. Appliquée à l'équation

$$
\left\{ \begin{array}{rcl} \dot{y} & = & f(y) \\ y(0) & = & y_0 \end{array} \right. \tag{3.2.1}
$$

la méthode $\mathcal{R}$ conduit à la récurrence

$$
\left\{ \begin{array}{rcl} y_{n+1} & = & y_n + h \sum_j b_j\, f(Y_j), \\ Y_i & = & y_n + h \sum_j a_{ij}\, f(Y_j),\;\; i = 1, \ldots, s. \end{array} \right. \tag{3.2.2}
$$

En différentiant cette relation par rapport à y_0, il vient alors :

$$
\left\{ \begin{array}{rcl} \partial_{y_0} y_{n+1} & = & \partial_{y_0} y_n + h \sum_j b_j\, f'(Y_j)\, \partial_{y_0} Y_j \\ \partial_{y_0} Y_i & = & \partial_{y_0} y_n + h \sum_j a_{i,j}\, f'(Y_j)\, \partial_{y_0} Y_j,\;\; i = 1, \ldots, s \end{array} \right. \tag{3.2.3}
$$

De même, $\mathcal{R}$ appliquée au système

$$
\left\{ \begin{array}{rcl} \dot{y} & = & f(y) \\ \dot{\Psi} & = & f'(y)\Psi \\ y(0) & = & y_0 \\ \Psi(0) & = & I_{\mathbb{R}^m} \end{array} \right. \tag{3.2.4}
$$

donne :

$$
\left\{ \begin{array}{rcl} y_{n+1} & = & y_n + h \sum_j b_j\, f(Y_j) \\ \Psi_{n+1} & = & \Psi_n + h \sum_j b_j\, f'(Y_j)\, \Psi_{n,j} \\ Y_i & = & y_n + h \sum_j a_{ij}\, f(Y_j),\;\; i = 1, \ldots, s \\ \Psi_{n,i} & = & \Psi_n + h \sum_j b_j\, f'(Y_j)\, \Psi_{n,j},\;\; i = 1, \ldots, s \end{array} \right. \tag{3.2.5}
$$

Donc si $\Psi_n = \partial_{y_0} y_n$, les équations coincident et $\Psi_{n+1} = \partial_{y_0} y_{n+1}$. On conclut en remarquant que $\Psi(0) = I_{\mathbb{R}^m} = \partial_{y_0} y_0$. Les calculs pour une methode partitionnée sont identiques. □

Lemme 3 *Une méthode de Runge-Kutta partitionnée de coefficients A, $\hat{A}$, b, $\hat{b}$ appliquée au système (3.1.1) conserve les invariants quadratiques, i.e. les invariants de la forme*

$$
Q(p, q) = p^T C q,
$$

si les conditions suivantes

$$B\hat{A}^T + A\hat{B} - b\hat{b}^T = 0,$$
$$b = \hat{b},$$

où $B = diag(b)$ et $\hat{B} = diag(\hat{b})$, sont satisfaites. Si en outre dans (3.1.1), f ne dépend que de q et g ne dépend que de p, alors la première condition est suffisante.

Preuve : En utilisant les équations (3.1.2), on obtient :

$$\begin{aligned} Q(p_1, q_1) &= \left(p_0 + h\sum_j b_j\, f(P_j, Q_j)\right)^T C\, \left(q_0 + h\sum_j \hat{b}_j\, g(P_j, Q_j)\right), \\ &= p_0^T C q_0 + h\sum_j \Big(b_j\, f^T(P_j, Q_j)\, C\, q_0 + \hat{b}_j\, p_0^T\, C\, g(P_j, Q_j)\Big). \\ &\quad +h^2 \sum_{j,\,k} b_j\, \hat{b}_k\, f^T(P_j, Q_j)\, C\, g(P_k, Q_k). \end{aligned}$$

Or

$$\begin{aligned} p_0 &= P_j - h\sum_k a_{jk}\, f(P_k, Q_k),\ j = 1, \ldots, s, \\ q_0 &= Q_j - h\sum_k \hat{a}_{jk}\, g(P_k, Q_k),\ j = 1, \ldots, s, \end{aligned}$$

donc

$$\begin{aligned} Q(p_1, q_1) &= p_0^T C q_0 + h\sum_j \underbrace{\Big(b_j\, f^T(P_j, Q_j)\, C\, Q_j + \hat{b}_j\, P_j^T\, C\, g(P_j, Q_j)\Big)}_{\text{terme nul car } f^T(p,q)Cq + p^T Cg(p,q) = 0} \\ &\quad -h^2 \sum_{j,\,k} \Big(b_j\, \hat{a}_{j\,k}\, f^T(P_j, Q_j)\, C\, g(P_k, Q_k) + \hat{b}_j\, a_{j\,k}\, f^T(P_k, Q_k)\, C\, g(P_j, Q_j)\Big) \\ &\quad +h^2 \sum_{j,\,k} \Big(b_j\, \hat{b}_k\, f^T(P_j, Q_j)\, C\, g(P_k, Q_k)\Big) \\ &= p_0^T C q_0 - h^2 \sum_{j,\,k} \underbrace{\Big(b_j\, \hat{a}_{j\,k} + \hat{b}_k\, a_{k\,j} - b_j\, \hat{b}_k\Big)}_{=0 \text{ par hypothèse}} f^T(P_j, Q_j)\, C\, g(P_k, Q_k) \end{aligned}$$

Finalement, $Q(p_1, q_1) = Q(p_0, q_0)$. Il reste à remarquer que si f ne dépend pas de p et si g ne dépend pas de q, alors l'invariance de $p^T Cq$ se traduit par

$$f^T(q)Cq + p^T Cg(p) = 0 \tag{3.2.6}$$

ce qui implique que $f^T(q)Cq = p^T Cg(p) = 0$. La seconde condition est alors superflue. □

On est maintenant en mesure d'énoncer et de démontrer partiellement la caractérisation suivante des méthodes de Runge-Kutta symplectiques.

Méthode	Ordre	Sympl.	Sym.	Expl.	Exemple
Gauss	$2s$	Oui	Oui	Non	Point milieu
PDIRK	$s = 1, 2, 3$	Oui	Oui	Oui	(3.3.1)

Théorème 7 *Soient A, $\hat{A}$, b, $\hat{b}$, les coefficients d'une méthode de Runge-Kutta partitionnée. Alors la méthode est symplectique si et seulement si*

$$B\hat{A}^T + A\hat{B} - b\hat{b}^T = 0,$$
$$b = \hat{b},$$

avec $B = diag(b)$, $\hat{B} = diag(\hat{b})$. Si l'hamiltonien est séparable, i.e. $H(p, q) = T(p) + V(q)$, la première condition est nécessaire et suffisante.

Preuve : Pour démontrer que la condition est nécessaire, il suffit d'interpréter la condition de symplecticité comme un invariant quadratique. Pour un système hamiltonien

$$\left\{ \begin{array}{rcl} \dot{y}(t) & = & J^{-1}\nabla H(y(t)) \\ y(0) & = & y_0 \end{array} \right. \tag{3.2.7}$$

on a en effet

$$\frac{d}{dt}(\Psi_t^T J \Psi_t) = 0 \tag{3.2.8}$$

où $\Psi_t = \frac{\partial \varphi_t(y_0)}{\partial y_0}$. Maintenant, si Φ_h est le flot numérique d'une méthode de Runge-Kutta,

$$(\Phi_h(y_0), \partial_{y_0}\Phi_h(y_0))$$

est, d'après le lemme 2, la solution numérique à l'instant h du système

$$\left\{ \begin{array}{rclrcl} \dot{y} & = & J^{-1}\nabla H(y), & y(0) & = & y_0 \\ \dot{\Psi}_t & = & J^{-1}(\nabla^2 H(y))\Psi_t, & \Psi(0) & = & I \end{array} \right.$$

pour lequel $\Psi_t^T J \Psi_t$ est un invariant quadratique. Si la méthode satisfait les conditions de symplecticité, cet invariant est donc conservé. Pour montrer que la condition énoncée est en outre nécessaire, il faut introduire une notion de méthode de Runge-Kutta "réduite", traduisant le fait que certaines étapes internes (autrement dit les Y_i) peuvent être redondantes. □

3.3 Quelques classes de méthodes symplectiques et/ou symétriques

A titre d'exemple, on présente ici une classe de méthodes dites PDIRK pour "Partitionned Diagonally Implicit Runge-Kutta", qui comme leur nom ne l'indique pas, sont *explicites* et symplectiques pour les systèmes Hamiltoniens séparés. Une autre classe de méthode très classique est celle des méthodes de Gauss, qui sont elles-aussi symplectiques à tout ordre.

Pour $b \in \mathbb{R}^s$ et $B \in \mathbb{R}^s$, on considère la méthode de Runge-Kutta partitionnée de matrices de coefficients

$$A = \begin{bmatrix} b_1 & 0 & 0 & \cdots & 0 \\ b_1 & b_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ b_1 & b_2 & b_3 & \cdots & b_s \end{bmatrix} \text{ et } \hat{A} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ B_1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ B_1 & B_2 & B_3 & \cdots & 0 \end{bmatrix}.$$

Alors l'implémentation pour un système séparable

$$\begin{aligned} \dot{p} &= f(q), \\ \dot{q} &= g(p), \end{aligned}$$

est explicite.

$$\begin{array}{llll} P_0 & = & p_n, & \\ Q_1 & = & q_n, & \\ for & i = 1, \cdots, s & do & \\ & P_i & = & P_{i-1} + hb_i f(Q_i), \\ & Q_{i+1} & = & Q_i + hB_i g(P_i), \\ end & & & \\ p_{n+1} & = & P_s, & \\ q_{n+1} & = & Q_{s+1}. & \end{array}$$

Pour

$$\begin{array}{rcl} b & = & (7/24, 3/4, -1/24)^T, \\ B & = & (2/3, -2/3, 1)^T \end{array} \tag{3.3.1}$$

on obtient une méthode symplectique d'ordre 3, explicite pour les systèmes séparables.

Chapitre 4

Analyse rétrograde

4.1 Idée générale

L'idée originale de l'analyse rétrograde est d'interpréter la solution numérique comme la solution exacte d'un problème approché.

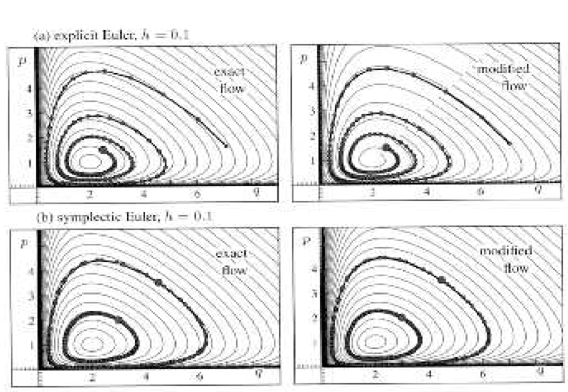
$$
\begin{array}{ccc}
 & & \varphi_t(y_0) \\
 & \nearrow & \\
\dot{y} = f(y) & & \\
 & \searrow & \\
 & & y_{n+1} = \Phi_h(y_n) \\
 & \nearrow & \\
\dot{\tilde{y}} = \tilde{f}^h(\tilde{y}) & &
\end{array}
$$

Le champ de vecteurs $\tilde{f}^h(\tilde{y})$ est déterminé sous la forme d'un développement en puissances de h en général non convergent ; l'équation différentielle modifiée (satisfaite par l'approximation numérique y_0, y_1, y_2, ...) est de la forme :

$$\dot{\tilde{y}} \;=\; f(\tilde{y}) + hf_2(\tilde{y}) + h^2 f_3(\tilde{y}) + \cdots\,,\ \tilde{y}(0) = y_0.$$

Résultat fondamental 1 : Sous certaines conditions portant sur f, et pour h suffisamment petit inférieur à un certain h^*, il existe une constante c et un N dépendant de h tels que

$$\forall\, nh\ \leq e^{\frac{h^*}{h}},\ \|\Phi_h^n(y_0) - \tilde{\varphi}_{N,\,nh}(y_0)\| \leq ce^{-\frac{h^*}{h}},$$



où $\tilde{\varphi}_{N,h}$ est le flot de l'équation tronquée au N-ième terme.

Résultat fondamental 2 : Si le système est *hamiltonien* et la méthode numérique *symplectique*, alors l'équation modifiée est encore *hamiltonienne*.

Conséquence : L'erreur numérique sur l'hamiltonien est en Ch^p si p est l'ordre de la méthode et ce pour des intervalles de temps exponentiellement longs.

4.2 Développement du flot numérique

Théorème 8 *On considère une méthode de Runge-Kutta de coefficients $A = (a_{i\,j}) \in \mathbb{R}^{s\times s}$ et $b = (b_i) \in \mathbb{R}^s$. Soient alors*

$$\mu = \sum_i |b_i| \text{ et } \kappa = \max_i \sum_j |a_{i,j}|. \tag{4.2.1}$$

Si le champ de vecteur f est supposé (réel-)analytique dans une boule complexe $B_{2R}(y_0)$ et est majoré par M sur $B_{2R}(y_0)$, alors pour h suffisamment petit, le flot numérique $\Phi_h(y)$ associé à la méthode de Runge-Kutta est analytique et les coefficients $d_j(y)$ de son développement

$$\Phi_h(y) = y + hf(y) + h^2 d_2(y) + \dots \tag{4.2.2}$$

sont eux-mêmes analytiques et satisfont :

$$\|d_j(y)\| \leq \mu M \left(\frac{2\kappa M}{R}\right)^{j-1} \text{ pour } y \in B_R(y_0). \tag{4.2.3}$$

Preuve : Soit $h_0 = \gamma \frac{R}{2\kappa M}$ pour $\gamma < 1$. On définit les fonctions $\Theta_i^{(n)}(h, y)$ par $\Theta_i^{(0)}(h, \cdot) \equiv Id_{\mathbb{C}^m}$ et pour $n \in \mathbb{N}$ par :

$$\begin{array}{rcl} \Theta_i^{(n+1)} \;:\; \{|h| \leq h_0\} \times B_R(y_0) & \longrightarrow & \mathbb{C}^m \\ (h, y) & \longmapsto & y + h \sum_j a_{i,j} f(\Theta_j^{(n)}(h, y)), \end{array}$$

m étant la dimension du système d'équations différentielles. Si pour tout j, tout $|h| \leq h_0$ et tout $y \in B_R(y_0)$, $\Theta_j^{(n)}(h, y) \in B_{3R/2}(y_0)$, alors il en est de même pour $\Theta_i^{(n+1)}$. En effet :

$$\max_{|h|\leq h_0,\, y \in B_R(y_0)} \|\Theta_i^{(n+1)}(h, y) - y\| \leq \max_{|h|\leq h_0,\, y\in B_R(y_0)} \|h \sum_j a_{i\,j} f(\Theta_j^{(n)}(h, y))\| \leq h_0\, \kappa\, M \leq \frac{R}{2}.$$

Les fonctions $\Theta_i^{(n)}(h, \cdot)$ prennent donc leurs valeurs dans $B_{3R/2}(y_0)$. En dérivant par rapport à z la fonction $\alpha(z) = f(y + z\Delta y)$ pour $y \in B_{3R/2}(y_0)$, $\|\Delta y\| \leq 1$ et $z \in B_{R/2}(0)$, une majoration de Cauchy donne

$$\|\alpha'(0)\| = \|f'(y)\Delta y\| \leq \frac{M}{R/2} = \frac{2M}{R}. \tag{4.2.4}$$

L'opérateur f' est donc borné par $(2M)/R$ sur $B_{3R/2}$ et il vient alors pour tous $m \geq n$:

$$\begin{aligned}
\max_{i,\,|h|\leq h_0,\, y\in B_R(y_0)} \|\Theta_i^{(m)}(h,y) - \Theta_i^{(n)}(h,y)\| &\leq \max_{i,\,|h|\leq h_0,\, y\in B_R(y_0)} \|h\sum_j a_{i,j}(f(\Theta_j^{(m-1)}(h,y)) - f(\Theta_j^{(n-1)}(h,y))\| \\
&\leq h_0\kappa\frac{2M}{R} \max_{j,\,|h|\leq h_0,\, y\in B_R(y_0)} \|\Theta_j^{(m-1)}(h,y) - \Theta_j^{(n-1)}(h,y)\| \\
&\dots \\
&\leq \gamma^n \max_{i,\,|h|\leq h_0,\, y\in B_R(y_0)} \|\Theta_i^{(m-n)}(h,y) - \Theta_i^{(0)}(h,y)\| \\
&\leq \gamma^n \frac{R}{2}
\end{aligned}$$

La suite des $\Theta_i^{(n)}$ converge donc uniformément vers une fonction Θ_i^∞ analytique (en h et y), puisque chacune des $\Theta_i^{(n)}$ est analytique. Finalement,

$$\|\Phi_h(h,y) - y\| \leq \mu M|h| \tag{4.2.5}$$

et

$$\|d_j(y)\| = \frac{1}{j!}\left\|\frac{d^j}{dh^j}(\Phi_h(y) - y)\bigg|_{h=0}\right\| \leq \frac{\mu\, M\, h_0}{h_0^j} \leq \mu M\left(\frac{2\kappa M}{\gamma R}\right)^{j-1}. \tag{4.2.6}$$

Le résultat s'obtient alors en faisant tendre γ vers 1. □

4.3 Développement de l'équation modifiée

4.3.1 Calcul des coefficients

Lemme 4 *Supposons que la méthode numérique Φ_h soit formellement développable en une série de la forme :*

$$\Phi_h(y) = y + hf(y) + h^2 d_2(y) + \dots \tag{4.3.1}$$

Alors, les coefficients f_j de l'équation modifiée

$$\frac{d}{dt}\tilde{y}^h(t) = f_1(\tilde{y}^h(t)) + hf_2(\tilde{y}^h(t)) + h^2 f_3(\tilde{y}^h(t)) + \dots \tag{4.3.2}$$

sont déterminés par la formule de récurrence suivante :

$$f_j(y) = d_j(y) - \sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1+\ldots+k_i=j, k_1\geq 1,\ldots,k_i\geq 1} (D_{k_1}\cdots D_{k_i} Id)(y), \tag{4.3.3}$$

où les D_i sont les opérateurs dérivée de Lie associés aux champs de vecteurs f_i.

Preuve : Soit $\tilde{\varphi}_t^h(y)$ le flot associé à l'équation modifiée :

$$\begin{aligned}
\frac{d}{dt}\tilde{y}^h(t) &= \tilde{f}^h(\tilde{y}^h(t)) &\quad (4.3.4)\\
&= f_1(\tilde{y}^h(t)) + hf_2(\tilde{y}^h(t)) + h^2 f_3(\tilde{y}^h(t)) + \dots &\quad (4.3.5)
\end{aligned}$$

Le développement en série formelle de $\tilde{\varphi}_t^h(y)$ s'écrit à partir de l'opérateur dérivée de Lie $\tilde{D}^h$ de $\tilde{f}^h$ sous la forme (cette expression est calculée en annexe) :

$$\tilde{\varphi}_t^h(y) = \sum_{i\geq 0} \frac{t^i}{i!}((\tilde{D}^h)^i Id)(y). \tag{4.3.6}$$

Or $\tilde{D}^h = D_1 + hD_2 + h^2 D_3 + \ldots$, et il vient donc :

$$\begin{aligned}
\tilde{\varphi}_t^h(y) &= \sum_{i\geq 0} \frac{t^i}{i!} \left(\sum_{k\geq 1} h^{k-1} D_k \right)^i (Id)(y) \\
&= y + \sum_{i\geq 1} \frac{t^i}{h^i i!} \sum_{k_1\geq 1,\ldots,k_i\geq 1} h^{k_1+\ldots+k_i} (D_{k_1}\cdots D_{k_i} Id)(y) \\
&= y + \sum_{j\geq 1} h^j \left(\sum_{i=1}^{j} \frac{t^i}{h^i i!} \sum_{k_1+\ldots+k_i=j, k_1\geq 1,\ldots,k_i\geq 1} (D_{k_1}\cdots D_{k_i} Id)(y) \right).
\end{aligned}$$

En identifiant les puissances de h, on a maintenant que $\tilde{\varphi}_h^h(y) = \Phi_h(y)$ si on a pour tout $j \geq 1$:

$$d_j(y) = \sum_{i=1}^{j} \frac{1}{i!} \sum_{k_1+\ldots+k_i=j, k_1\geq 1,\ldots,k_i\geq 1} (D_{k_1}\cdots D_{k_i} Id)(y) \tag{4.3.7}$$

Le terme de la somme correspondant à $i = 1$ n'est autre que $(D_j Id)(y) = f_j(y)$ et le résultat est alors immédiat. □

Pour les trois premiers termes du développement, on obtient :

$$\begin{aligned}
f_1(y) &= d_1(y) \\
&= f(y), \\
f_2(y) &= d_2(y) - \frac{1}{2}(D_1^2 Id)(y) \\
&= d_2(y) - \frac{1}{2}(D_1 f_1)(y) \\
&= d_2(y) - \frac{1}{2}\left(\frac{\partial f_1}{\partial y} f_1\right)(y), \\
f_3(y) &= d_3(y) - \frac{1}{2}(D_1 D_2 + D_2 D_1)(Id)(y) - \frac{1}{6}(D_1^3 Id)(y) \\
&= d_3(y) - \frac{1}{2}\left(\frac{\partial f_2}{\partial y} f_1 + \frac{\partial f_1}{\partial y} f_2\right)(y) - \frac{1}{6}\left(\frac{\partial^2 f_1}{\partial y^2}(f_1, f_1) + \frac{\partial f_1}{\partial y}\frac{\partial f_1}{\partial y} f_1\right)(y).
\end{aligned}$$

4.3.2 Majoration des coefficients

Lemme 5 (Lemme technique) *Soient $f : \mathbb{R}^n \to \mathbb{R}^n$ et $g : \mathbb{R}^n \to \mathbb{R}^m$ deux fonctions analytiques et soient σ et ρ deux réels tels que $0 \leq \sigma < \rho$. On désigne par $\|\cdot\|_r$ la norme sup sur $B_r(y_0)$ et par D_f la dérivée de Lie associée au champ f. Alors on a la majoration suivante :*

$$\|D_f g\|_\sigma \leq \frac{\|f\|_\sigma \|g\|_\rho}{\rho - \sigma} \tag{4.3.8}$$

Preuve : Soit, pour $y \in B_\sigma(y_0)$ et $|z| \le \varepsilon := \frac{\rho-\sigma}{\|f\|_\sigma}$, $\alpha(z) = g(y + zf(y))$. La majoration de Cauchy pour α donne alors :

$$\|(D_f g)(y)\| = \|g'(y)f(y)\| = \|\alpha'(0)\| \le \frac{\|g\|_\rho}{\varepsilon} = \frac{\|f\|_\sigma \|g\|_\rho}{\rho - \sigma}. \tag{4.3.9}$$

Théorème 9 *On suppose que le champ de vecteur f est analytique sur $B_{2R}(y_0)$ et borné par M sur cette boule. On suppose par ailleurs, que la majoration (4.2.3) des coefficients d_j du développement en série de Taylor de la méthode numérique est valable sur la boule $B_R(y_0)$. Alors les coefficients f_j de l'équation modifiée tels que calculés dans le lemme 4 sont majorés par :*

$$\forall\, y \in B_{R/2}(y_0),\ \|f_j(y)\| \le (\log 2)\eta M \left(\frac{\eta M j}{R}\right)^{j-1}, \tag{4.3.10}$$

où $\eta = 2\max(\kappa, \mu/(2\log 2 - 1))$.

Preuve : Soit $J \ge 2$ un entier fixé et $\delta = \frac{R}{2(J-1)}$. On définit, pour $j = 1, \ldots, J$, les rayons

$$r_j = \left(\frac{j-1}{J-1}\right)\frac{R}{2} + \left(\frac{J-j}{J-1}\right) R \tag{4.3.11}$$

de sorte que $r_1 = R$ et $r_J = R/2$. Pour un ensemble de i indices $k_1, \ldots, k_i$, supposés non nuls et de somme j, on estime alors le terme général de (4.3.3) par l'utilisation répétée du lemme précédent :

$$\begin{aligned} \|D_{k_1} \cdots D_{k_i} Id\|_{r_j} &\le \frac{1}{r_{j-1} - r_j} \|f_{k_1}\|_{r_j} \|D_{k_2} \cdots D_{k_i} Id\|_{r_{j-1}} \\ &\le \cdots \\ &\le \frac{1}{\delta^{i-1}} \prod_{l=1}^{i} \|f_{k_l}\|_{r_{j-l+1}} \end{aligned}$$

Or, $k_l \le j - i + 1 \le j - l + 1$, donc $r_{k_l} \ge r_{j-l+1}$ et $\|f_{k_l}\|_{r_{j-l+1}} \le \|f_{k_l}\|_{r_{k_l}}$. On a donc :

$$\|D_{k_1} \cdots D_{k_i} Id\|_{r_j} \le \frac{1}{\delta^{i-1}} \prod_{l=1}^{i} \|f_{k_l}\|_{r_{k_l}}. \tag{4.3.12}$$

En insérant cette majoration dans l'expression de f_j calculée dans le lemme 4, on obtient alors, pour $1 \le j \le J$:

$$\|f_j\|_{r_j} \le \|d_j\|_R + \sum_{i=2}^{j} \frac{1}{\delta^{i-1}\, i!} \sum_{k_1+\ldots+k_i=j} \prod_{l=1}^{i} \|f_{k_l}\|_{r_{k_l}}. \tag{4.3.13}$$

En définissant β_j pour $j \ge 1$ par :

$$\beta_j = \frac{\mu M}{\delta}\left(\frac{2\kappa M}{R}\right)^{j-1} + \sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1+\ldots+k_i=j} \prod_{l=1}^{i} \beta_{k_l}, \tag{4.3.14}$$

on peut montrer par une récurrence facile, que[1] $\|f_j\|_{r_j} \le \delta\beta_j$. Afin d'estimer les β_j, on définit maintenant la fonction génératrice

$$b(\zeta) = \sum_{j \ge 1} \beta_j \zeta^j. \tag{4.3.15}$$

[1]Notons au passage que $\beta_1 = \frac{\mu M}{\delta}$ et que $\|f_1\|_{r_1} = \|f\|_R \le \|f\|_{2R} \le M$.

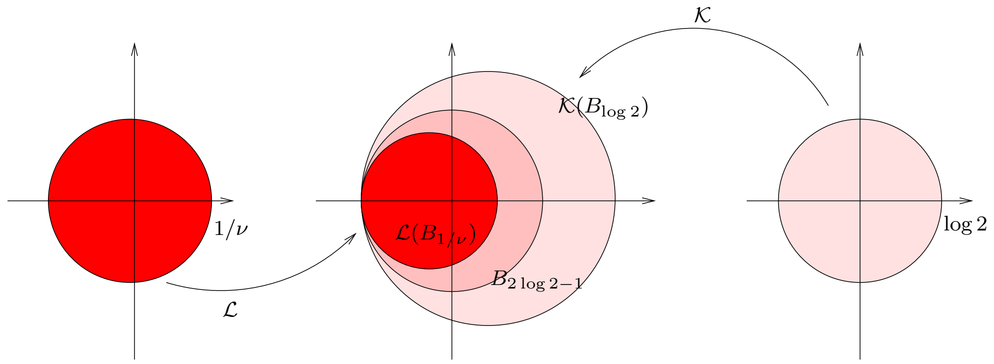


FIG. 4.1 – Intersection des images de $\mathcal{K}$ et $\mathcal{L}$

De (4.3.14), on tire :

$$\beta_j\zeta^j = \frac{\mu M}{\delta}\zeta\left(\frac{2\kappa M}{R}\zeta\right)^{j-1} + \sum_{i=2}^{j}\frac{1}{i!}\sum_{k_1+\ldots+k_i=j}\prod_{l=1}^{i}\beta_{k_l}\zeta^{k_l},$$

puis

$$\sum_{j\geq 1}\beta_j\zeta^j = \frac{\mu M}{\delta}\zeta\sum_{j\geq 1}\left(\frac{2\kappa M}{R}\zeta\right)^{j-1} + \sum_{j\geq 2}\sum_{i=2}^{j}\frac{1}{i!}\sum_{k_1+\ldots+k_i=j}\prod_{l=1}^{i}\beta_{k_l}\zeta^{k_l},$$

c'est-à-dire :

$$\begin{aligned} b(\zeta) &= \frac{\mu M}{\delta}\zeta\frac{1}{1-\frac{2\kappa M}{R}\zeta} + \sum_{i\geq 2}\frac{1}{i!}\sum_{j\geq i}\sum_{k_1+\ldots+k_i=j}\prod_{l=1}^{i}\beta_{k_l}\zeta^{k_l}, \\ &= \frac{\mu M}{\delta}\zeta\frac{1}{1-\frac{2\kappa M}{R}\zeta} + \sum_{i\geq 2}\frac{1}{i!}\sum_{k_1,\ldots,k_i}\prod_{l=1}^{i}\beta_{k_l}\zeta^{k_l}, \\ &= \frac{\mu M}{\delta}\zeta\frac{1}{1-\frac{2\kappa M}{R}\zeta} + \sum_{i\geq 2}\frac{1}{i!}b(\zeta)^i. \end{aligned}$$

Finalement, en notant $\gamma = \mu M/\delta$ et $q = 2\kappa M/R$, on obtient la relation suivante :

$$b(\zeta) = \frac{\gamma\zeta}{1-q\zeta} + e^{b(\zeta)} - 1 - b(\zeta). \tag{4.3.16}$$

Soient alors les applications :

$$\mathcal{K}\ : b \to e^b - 1 - 2b \text{ et } \mathcal{L}\ : \zeta \to \frac{-\gamma\zeta}{1-q\zeta} \tag{4.3.17}$$

L'image par $\mathcal{K}$ de la boule de centre $b=0$ et de rayon $\log 2$ contient la boule de centre 0, $B_{2\log 2-1}$. Par ailleurs, l'image par $\mathcal{L}$ de la boule de centre 0 et de rayon $1/\nu$ avec $1/\nu = \frac{2\log 2-1}{\gamma+q(2\log 2-1)}$ est contenue dans $B_{2\log 2-1}$, et ce, quels que soient $\gamma \geq 0$ et $q \geq 0$. On a ainsi (voir figure 4.1) :

$$\mathcal{L}(B_{1/\nu}) \subset B_{2\log 2-1} \subset \mathcal{K}(B_{\log 2}). \tag{4.3.18}$$

D'autre part, pour $|b| < \log 2$, $\mathcal{K}'(b) \neq 0$ et donc la fonction $\mathcal{K}$ est localement inversible sur $B_{2\log 2-1}$, et son inverse est analytique. On peut alors l'étendre sur tout $B_{2\log 2-1}$ et donc en particulier sur $\mathcal{L}(B_{1/\nu})$. En conclusion, la fonction $b(\zeta) = (\mathcal{K}^{-1} \circ \mathcal{L})(\zeta)$ est analytique sur $B_{1/\nu}$. En outre, on a :

$$\forall \zeta \in B_{1/\nu},\ |b(\zeta)| < \log 2, \tag{4.3.19}$$

ce qui implique finalement,

$$\|f_J\|_{r_J} = \|f_J\|_{R/2} \le \delta\beta_J \le \log 2\, \delta\, \nu^J. \tag{4.3.20}$$

On obtient le résultat en remarquant que

$$\begin{aligned} \nu &= q + \frac{\gamma}{2\log 2 - 1} = \frac{2\kappa M}{R} + \frac{\mu M}{\delta(2\log 2 - 1)} = \frac{M}{R}\left(2\kappa + 2\frac{\mu(J-1)}{2\log 2 - 1}\right) \\ &\le 2\max(\kappa, \mu/(2\log 2 - 1))\frac{M\,J}{R}. \end{aligned}$$

Soit

$$\begin{aligned} \|f_J\|_{R/2} &\le \log 2\, \frac{R}{2(J-1)} \left(\frac{\eta M J}{R}\right)^J \\ &\le \log 2\, \eta\, M \left(\frac{\eta M J}{R}\right)^{J-1}. \end{aligned}$$

□

4.4 Propriétés de l'équation modifiée

Théorème 10 *Soit Φ_h le flot numérique associé à une méthode de Runge-Kutta d'ordre p et $\delta_{p+1}(y)$ le terme d'erreur principal :*

$$\Phi_h(y) = \varphi_h(y) + h^{p+1}\delta_{p+1}(y) + \mathcal{O}(h^{p+2}), \tag{4.4.1}$$

où $\varphi_t(y)$ désigne le flot exact de l'équation différentielle $\dot{y} = f(y)$. Alors l'équation modifiée est de la forme :

$$\begin{aligned} \dot{\tilde{y}} &= f(\tilde{y}) + h^p\delta_{p+1}(y) + h^{p+1}f_{p+2}(\tilde{y}) + \dots \\ \tilde{y}_0 &= y_0. \end{aligned}$$

Preuve : C'est une conséquence immédiate de la définition des coefficients de l'équation modifiée. On a en effet :

$$\tilde{\varphi}_h^h(y) = \Phi_h(y) = \varphi_h(y) + h^{p+1}\delta_{p+1}(y) + \mathcal{O}(h^{p+2})$$

et donc :

$$\sum_{j\ge 1} h^j \left(\sum_{i=1}^{j} \frac{1}{i!} \sum_{k_1+\ldots+k_i=j, k_1\ge 1,\ldots,k_i\ge 1} (D_{k_1}\cdots D_{k_i}Id)(y)\right) = \sum_{j\ge 1}\frac{h^j}{j!}(D^jId)(y) + h^{p+1}\delta_{p+1}(y) + \mathcal{O}(h^{p+2})$$

On obtient alors successivement :

$$\begin{array}{ll}
j=1\ : & f_1 \equiv f \\
j=2\ : & f_2 + \frac{1}{2!}(D_1 f_1) = \frac{1}{2!}(Df) \Rightarrow f_2 \equiv 0 \\
\ldots & \ldots \\
j=p\ : & f_p + \frac{1}{p!}(D_1^{p-1} f_1) = \frac{1}{2!}(D^{p-1} f) \Rightarrow f_p \equiv 0 \\
j=p+1\ : & f_{p+1} + \frac{1}{(p+1)!}(D_1^p f_1) = \frac{1}{(p+1)!}(D^p f) + \delta_{p+1} \Rightarrow f_{p+1} \equiv \delta_{p+1}.
\end{array}$$

□

Théorème 11 *Soit Φ_h le flot numérique associé à une méthode de Runge-Kutta symplectique appliquée à un système Hamiltonien d'Hamiltonien H de classe C^∞ sur $\mathbb{R}^{2d}$. Alors l'équation modifiée est elle-aussi Hamiltonienne, c'est-à-dire que pour tout $j \in \mathbb{N}^*$, il existe un Hamiltonien H_j de classe C^∞ sur $\mathbb{R}^{2d}$ tel que*

$$f_j(y) = J^{-1}\nabla H_j(y). \tag{4.4.2}$$

Preuve : La preuve se fait par récurrence. Pour $j = 1$, $f_1 \equiv f$ et le résultat est vrai. Supposons maintenant que pour $r \geq 1$, la relation (4.4.2) soit établie pour tout $1 \leq j \leq r$. Le flot $\tilde{\varphi}_{r,h}(y)$ correspondant à l'équation modifiée

$$\dot{\tilde{y}} \quad = \quad f(\tilde{y}) + hf_1(y) + \ldots + h^{r-1} f_r(y) \tag{4.4.3}$$

est le flot d'un système Hamiltonien, donc est symplectique. Or :

$$\Phi_h(y) = \tilde{\varphi}_{r,h}(y) + h^{r+1} f_{r+1}(y) + \mathcal{O}(h^{r+2}) \tag{4.4.4}$$

donc

$$\Phi_h'(y) = (\tilde{\varphi}_{r,h})'(y) + h^{r+1} f_{r+1}'(y) + \mathcal{O}(h^{r+2}). \tag{4.4.5}$$

Les flots $\Phi_h(y)$ et $\tilde{\varphi}_{r,h}(y)$ étant symplectiques, on a alors :

$$\begin{array}{rcl}
J = \left(\Phi_h'(y)\right)^T J\Phi_h'(y) & = & \left(\tilde{\varphi}_{r,h})'(y)\right)^T J\tilde{\varphi}_{r,h})'(y) \\
& & +h^{r+1}\left[\left(\tilde{\varphi}_{r,h})'(y)\right)^T Jf_{r+1}'(y) + \left(f_{r+1}'(y)\right)^T J(\tilde{\varphi}_{r,h})'(y)\right] \\
& & +\mathcal{O}(h^{r+2}) \\
& \overset{(\tilde{\varphi}_{r,h})'(y)=I+\mathcal{O}(h)}{=} & J + h^{r+1}\left[Jf_{r+1}'(y) + \left(f_{r+1}'(y)\right)^T J\right] + \mathcal{O}(h^{r+2}).
\end{array}$$

En conséquence, la matrice $Jf_{r+1}'(y)$ est symétrique et d'après le lemme d'intégrabilité, on a :

$$Jf_{r+1}'(y) = \nabla H_{r+1}(y), \tag{4.4.6}$$

où H est défini sur $\mathbb{R}^{2d}$ et de classe C^∞. □

4.5 Conservation de l'énergie

En tronquant l'équation modifiée de manière optimale, on peut maintenant montrer la conservation de l'Hamiltonien sur des temps exponentiellement longs en l'inverse du pas d'intégration.

4.5.1 Troncature optimale de l'équation modifiée

Soit donc, pour $N \geq 2$, l'équation modifiée tronquée au N-ième terme :

$$\dot{\tilde{y}} = f(\tilde{y}) + hf_2(\tilde{y}) + \ldots + h^{N-1} f_N(\tilde{y}), \tag{4.5.1}$$

et soit $\tilde{\varphi}_{N,h}(y)$ le flot exact correspondant. On souhaite ici déterminer s'il existe un N minimisant l'erreur entre le flot numérique Φ_h et le flot $\tilde{\varphi}_{h,N}(y)$. Notons tout d'abord que si Φ_h admet un développement de la forme (4.2.2) et que les termes du développement sont majorés comme en (4.2.3), alors le dit développement est convergent pour $|h| \leq \frac{R}{2\kappa M}$. Symétriquement, le flot $\tilde{\varphi}_{N,h}(y)$ se développe en puissance de h et de l'opérateur $\tilde{D}_h = D + hD_1 + \ldots h^{N-1} D_N$:

$$\begin{aligned} \tilde{\varphi}_{h,N} &= Id + \sum_{i \geq 1} \frac{h^i}{i!} \left(\sum_{k=1}^{N} h^{k-1} D_k \right)^i (Id) \\ &= Id + \sum_{j \geq 1} h^j G_j Id \end{aligned}$$

où les G_j sont les opérateurs

$$G_j = \sum_{i=1}^{j} \frac{1}{i!} \sum_{1 \leq k_1, \ldots, k_i \leq N,\ k_1 + \ldots + k_i = j} (D_{k_1} \cdots D_{k_i}). \tag{4.5.2}$$

Afin de majorer les $G_j Id$ sur la boule $B_{R/4}(y_0)$, on procède comme dans le théorème 9, avec $\delta = R/(4(i-1))$. Pour $i \geq 2$ indices k_l, $1 \leq k_l \leq N$ de somme j, on a :

$$\begin{aligned} \|D_{k_1} \cdots D_{k_i} Id\|_{R/4} &\leq \frac{1}{\delta^{i-1}} \prod_{l=1}^{i} \|f_{k_l}\|_{R/2} \\ &\leq (\eta M \log 2) \left(\frac{4(i-1)\log 2}{N} \right)^{i-1} \left(\frac{\eta M N}{R} \right)^{j-1}. \end{aligned}$$

Il vient alors :

$$\|G_j(Id)\|_{R/4} \leq (\eta M \log 2) \left(\frac{\eta M N}{R} \right)^{j-1} \sum_{i=1}^{j} \frac{1}{i!} \left(\frac{4(i-1)\log 2}{N} \right)^{i-1} \mathrm{Card} S(i,j,N) \tag{4.5.3}$$

où $S(i,j,N) = \{(k_1, \ldots, k_i) \in [1,N]^i, k_1 + \ldots + k_i = j\}$, de sorte que $\mathrm{Card}S(i,j,N) \leq N^{i-1}$ et donc :

$$\|G_j(Id)\|_{R/4} \leq (\eta M \log 2) \left(\frac{\eta M N}{R} \right)^{j-1} \sum_{i=1}^{j} \frac{(4(j-1)\log 2)^{i-1}}{i!} \leq (\eta M \log 2) \left(\frac{16 \eta M N}{R} \right)^{j-1} \tag{4.5.4}$$

En particulier, le développement de $\tilde{\varphi}_{N,h}$ est convergent pour $y \in B_{R/4}(y_0)$ et $|hN| \leq \frac{R}{16\eta M}$.

On est maintenant en mesure d'énoncer le résultat principal de l'analyse rétrograde :

Théorème 12 *On suppose que le champ de vecteur f est analytique sur $B_{2R}(y_0)$ et borné par M sur cette boule. On suppose par ailleurs, que la majoration (4.2.3) des coefficients d_j du développement en série de Taylor de la méthode numérique Φ_h est valable sur la boule $B_R(y_0)$. Soit $h_0 = \frac{R}{16\eta Me}$: alors, si $|h| \leq h_0$, on a pour $N = E(h_0/h)$ la majoration suivante :*

$$\forall\, y \in B_{R/4}(y_0),\ \|\Phi_h(y) - \tilde{\varphi}_{N,h}(y)\| \leq h\gamma M e^{-\frac{h_0}{h}}, \tag{4.5.5}$$

où $\gamma = \frac{(2e^2\eta \log 2)}{e-1}$.

Preuve : Les champs de vecteurs f_j, $1 \leq j \leq N$ sont construits de manière à ce que les termes des développements de Φ_h et $\tilde{\varphi}_{N,h}$ coincident jusqu'à l'ordre N. On a donc :

$$\Phi_h - \tilde{\varphi}_{N,h} = \sum_{j \geq N+1} h^j \left(d_j - G_j Id\right) \tag{4.5.6}$$

Ainsi :

$$\begin{aligned}
\|\Phi_h - \tilde{\varphi}_{N,h}\|_{R/4} &\leq \sum_{j \geq N+1} h^j \, \|d_j - G_j(Id)\|_{R/4} \\
&\leq \sum_{j \geq N+1} h^j \left(\mu M \left(\frac{2\kappa M}{R}\right)^{j-1} + (\eta M \log 2) \left(\frac{16 \eta M N}{R}\right)^{j-1} \right) \\
&\leq (2\eta M \log 2)\, h \sum_{j \geq N} \left(\frac{16 \eta M N h}{R}\right)^j
\end{aligned}$$

Pour $Nh \leq h_0$, on a alors :

$$\|\Phi_h - \tilde{\varphi}_{N,h}\|_{R/4} \leq (2\eta M \log 2)\, h \left(\frac{hN}{eh_0}\right)^N \frac{e}{e-1} \leq \frac{(2e\eta M \log 2)}{e-1}\, h\, e^{-N} \tag{4.5.7}$$

car la fonction $x \mapsto (\epsilon x)^x$ admet pour $\epsilon < 1$ un minimum en $x = (\epsilon e)^{-1}$, et pour $0 < x \leq (\epsilon e)^{-1}$, $(\epsilon x)^x \leq e^{-x}$ (ici, $\epsilon = \frac{h}{h_0 e}$). Pour $N = E(h_0/h)$ on obtient finalement :

$$\|\Phi_h - \tilde{\varphi}_{N,h}\|_{R/4} \leq \frac{(2e^2 \eta M \log 2)}{e-1}\, h\, e^{-\frac{h_0}{h}}.$$

□

4.5.2 Conservation de l'énergie sur des temps "longs"

Dans le cas d'une système Hamiltonien, on a vu (Théorème 11) que les termes f_j de l'équation modifiée pouvaient s'écrire $f_j(y) = J^{-1}\nabla_y H(y)$. L'équation modifiée tronquée au N-ième terme est donc une équation Hamiltonienne correspondant à l'Hamiltonien

$$\tilde{H} = H + h^p H_{p+1} + \dots h^{N-1} H_N \tag{4.5.8}$$

que l'on suppose défini sur le même ensemble que H (c'est le cas en particulier si H est défini sur un ouvert étoilé). Dans les conditions du théorème précédent, on a alors :

Théorème 13 *On considère un Hamiltonien analytique $H \ : D \subset \mathbb{R}^{2d} \to \mathbb{R}$ où D est un ouvert étoilé et une méthode numérique Φ_h vérifiant les inégalités (4.2.3) appliquée au système différentiel d'Hamiltonien H. Si la solution numérique $y_n = \Phi_h^n(y_0)$, $n = 0, \dots,$ reste dans un compact $K \subset D$, alors il existe $h_0 > 0$ tel que :*

$$\begin{aligned}
\tilde{H}(y_n) &= \tilde{H}(y_0) + \mathcal{O}(e^{-\frac{h_0}{2h}}), && (4.5.9)\\
H(y_n) &= H(y_0) + \mathcal{O}(h^p), && (4.5.10)
\end{aligned}$$

sur des intervalles de temps $nh \leq e^{\frac{h_0}{2h}}$.

Preuve : K étant compact et D ouvert, il existe $R > 0$ tel que pour tout $y \in K$, $B_R(y) \subset D$ et on peut majorer uniformément $f = \nabla H$, par M, sur le compact $\{z \in D; \exists\, y \in K, \|y - z\| \leq R\}$. Toutes les majorations obtenues sucessivement dans les théorèmes 4.2.3, 9 et 12 sont donc applicables uniformément sur K avec un h_0 et un $N = E(h_0/h)$ uniformes eux-aussi. Maintenant,

$$\begin{aligned} \tilde{H}(y_n) - \tilde{H}(y_0) &= \sum_{j=1}^{n} \Big(\tilde{H}(y_j) - \tilde{H}(y_{j-1})\Big) = \sum_{j=1}^{n} \Big(\tilde{H}(y_j) - \tilde{H}(\tilde{\varphi}_{N,h}(y_{j-1}))\Big) && (4.5.11) \\ &= \sum_{j=1}^{n} \Big(\tilde{H}(\Phi_h(y_{j-1})) - \tilde{H}(\tilde{\varphi}_{N,h}(y_{j-1}))\Big) && (4.5.12) \end{aligned}$$

car le flot $\tilde{\varphi}_{N,h}$ conserve $\tilde{H}$. De plus, $\nabla\tilde{H} = \nabla H + \ldots + h^{N-1}\nabla H_{N+1} = J(f + \ldots + h^{N-1} f_N)$ où les f_j sont majorés sur K par (9) d'où il résulte que $\tilde{H}$ est L-Lipschitz uniformément en $|h| \leq h_0$ et donc :

$$\|\tilde{H}(y_n) - \tilde{H}(y_0)\| \leq L \sum_{j=1}^{n} \|\Phi_h(y_{j-1}) - \tilde{\varphi}_{N,h}(y_{j-1}))\| \leq LM\gamma e^{-\frac{h_0}{h}} nh \leq LM\gamma e^{-\frac{h_0}{2h}} \qquad (4.5.13)$$

On conclut en remarquant que $H(y_n) = \tilde{H}(y_n) + \mathcal{O}(h^p)$. $\square$

Chapitre 5

Systèmes hamiltoniens complètement intégrables

L'espoir des mathématiciens des $18^{\text{ième}}$ et $19^{\text{ième}}$ siècles était de transformer un système d'Hamiltonien $H(p,q)$ quelconque en un système plus simple, intégrable "à la main", ce qui est en particulier le cas si, après le changement de variables $(p,q) \to (x,y)$, le système est encore Hamiltonien pour $K(x) = H(p,q)$. En effet, si une telle transformation existe, alors les équations en (x,y) s'écrivent :

$$\begin{aligned} \dot{x} &= 0, & (5.0.1)\\ \dot{y} &= (\nabla_x K)(x), & (5.0.2) \end{aligned}$$

et s'intègrent immédiatement en :

$$\begin{aligned} x(t) &= x_0, & (5.0.3)\\ y(t) &= y_0 + t(\nabla_x K)(x_0). & (5.0.4) \end{aligned}$$

Pour que le système en (x,y) soit Hamiltonien, il est nécessaire (et suffisant cf Théorème 6) que le changement de variables $(p,q) \to (x,y)$ soit symplectique. Il existe principalement deux manières de construire des difféomorphismes symplectiques : soit en les exprimant comme le flot d'un système Hamiltonien, soit en les définissant à partir d'une fonction scalaire S dite *fonction génératrice*. La seconde technique fait l'objet du lemme suivant.

Lemme 6 *Soit $S(x,q)$ une fonction de $C^2(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$ telle que la matrice Hessienne $\nabla^2_{x,q} S$ soit inversible au voisinage du point (x_0, q_0). Alors les équations*

$$\begin{aligned} y &= \nabla_x S(x,q), & (5.0.5)\\ p &= \nabla_q S(x,q), & (5.0.6) \end{aligned}$$

définissent un difféomorphisme $(x,y) = l(p,q) = (F(p,q), G(p,q))$ d'un voisinage de (p_0,q_0) dans un voisinage de (x_0,y_0) où $p_0 = \nabla_q S(x_0,q_0)$ et $y_0 = \nabla_x S(x_0,q_0)$. De plus, cette transformation est symplectique.

Preuve : Le théorème des fonctions implicites permet d'affirmer qu'il existe F de $\mathbb{R}^{2d}$ dans $\mathbb{R}^d$ régulière telle que $\nabla_q S(F(p,q),q) = p$. Alors $l(p,q) = (F(p,q), \nabla_x S(F(p,q),q))$ existe et on a :

$$\begin{aligned} \partial_p F(p,q) &= (\nabla^2_{q,x} S)^{-1}(F,q), & (5.0.7)\\ \partial_q F(p,q) &= -(\nabla^2_{q,x} S)^{-1}(F,q) \cdot (\nabla^2_{q,q} S)(F,q). & (5.0.8) \end{aligned}$$

Il vient donc :

$$l'(p,q) = \left[\begin{array}{cc} \partial_p F & \partial_q F \\ (\nabla^2_{x,x} S)\partial_p F & (\nabla^2_{x,x} S)\partial_q F + (\nabla^2_{x,q} S) \end{array} \right]. \tag{5.0.9}$$

Ainsi, $l'^T J l' = J$ si et seulement si

$$\begin{array}{rcl} -(\partial_p F)^T (\nabla^2_{x,x} S)(\partial_p F) + (\partial_p F)^T (\nabla^2_{x,x} S)(\partial_p F) & = & 0, \\ -(\partial_p F)^T (\nabla^2_{x,x} S)(\partial_q F) + (\partial_p F)^T (\nabla^2_{x,x} S)(\partial_q F) + (\partial_p F)^T (\nabla^2_{x,q} S) & = & I_{\mathbb{R}^d}, \\ -(\partial_q F)^T (\nabla^2_{x,x} S)(\partial_p F) + (\partial_q F)^T (\nabla^2_{x,x} S)(\partial_p F) - (\nabla^2_{q,x} S)(\partial_p F) & = & -I_{\mathbb{R}^d}, \\ -(\partial_q F)^T (\nabla^2_{x,x} S)(\partial_q F) + (\partial_q F)^T (\nabla^2_{x,x} S)(\partial_q F) - (\nabla^2_{q,x} S)(\partial_q F) + (\partial_q F)^T (\nabla^2_{x,q} S) & = & 0. \end{array}$$

Il suffit alors de remarquer que $\nabla^2_{x,x} S$ est symétrique, que $(\nabla^2_{q,x} S)^T = \nabla^2_{x,q} S$ et d'insérer les expressions obtenues en (5.0.7) et (5.0.8). $\square$

Si la fonction S est telle que $H(l^{-1}(x,y)) = K(x)$, alors, d'après (5.0.1), les composantes $x_i = F_i(p,q)$ de F sont des intégrales premières du système d'Hamiltonien $H(p,q)$. En outre, l'expression obtenue pour

$$(\partial_p F)(\partial_q F)^T = -(\nabla^2_{q,x} S)^{-1} (\nabla^2_{q,q} S)(\nabla^2_{q,x} S)^{-T} = (\partial_q F)(\partial_p F)^T$$

montre que

$$[\partial_p F, \partial_q F] J^{-1} \left[\begin{array}{c} (\partial_p F)^T \\ (\partial_q F)^T \end{array} \right] = 0, \tag{5.0.10}$$

de sorte que les F_i sont en *involution* deux à deux, c'est-à-dire que pour $i = 1, \ldots, d$ et $j = 1, \ldots, d$,

$$\{F_i, F_j\} = (\nabla_{(p,q)} F_i)^T J^{-1} (\nabla_{(p,q)} F_j) = 0.$$

Réciproquement, on a le lemme suivant :

Lemme 7 (Liouville) *Soit F une fonction définie dans un voisinage de $(p_0, q_0) \in \mathbb{R}^d \times \mathbb{R}^d$ à valeurs dans $\mathbb{R}^d$. On suppose que les composantes F_i de F sont en involution deux à deux, c'est-à-dire sont telles que pour $i, j = 1, \ldots, d$,*

$$\{F_i, F_j\} = 0, \tag{5.0.11}$$

et qu'en outre $\frac{\partial F}{\partial(p,q)}(p_0, q_0)$ est de rang d. Alors, il existe une fonction G définie dans un voisinage de (p_0, q_0) à valeurs dans $\mathbb{R}^d$ telle que

$$l \ : (p,q) \to (x,y) = (F(p,q), G(p,q)) \tag{5.0.12}$$

définisse un difféomorphisme symplectique l au voisinage de (p_0, q_0).

Remarque 4 *Le système d'Hamiltonien $H_v(p,q)$ devient, par le changement de variables* **local** *l, un système d'Hamiltonien $K(x,y) = (H_v \circ l^{-1})(x,y)$. Or*

$$H_v(l^{-1}(x,y)) = \sum_i v_i F_i(l^{-1}(x,y)) = \sum_i v_i x_i = v^T x. \tag{5.0.13}$$

Donc K ne dépend que de x, et le système différentiel associé à K s'écrit donc :

$$\begin{array}{rcl} \dot{x} & = & -\nabla_y K(x) = 0 \\ \dot{y} & = & \nabla_x K(x) \end{array} \begin{array}{r} (5.0.14) \\ (5.0.15) \end{array}$$

et est donc bien de la forme recherchée.

Preuve : Quitte à effectuer quelques tranformations symplectiques linéaires sur $[\partial_p F, \partial_q F]$, on peut supposer que $\partial_p F$ est inversible. Comme précédemment, le théorème des fonctions implicites permet d'affirmer l'existence d'une fonction $p = P(x,q)$ telle que $F(P(x,q),q) = x$, définie au voisinage de (x_0, q_0), où $x_0 = F(p_0, q_0)$. On a alors :

$$\partial_x P = (\partial_p F)^{-1} \text{ et } \partial_q P = -(\partial_p F)^{-1}(\partial_q F). \tag{5.0.16}$$

La condition d'involution (confère (5.0.10)) implique que :

$$(\partial_p F)(\partial_q F)^T = (\partial_q F)(\partial_p F)^T \tag{5.0.17}$$

i.e., après multiplication à gauche par $(\partial_p F)^{-1}$ et à droite par $(\partial_p F)^{-T}$:

$$(\partial_q P) = (\partial_q P)^T. \tag{5.0.18}$$

D'après le lemme d'intégrabilité, on peut donc écrire localement $P(x,q) = (\nabla_q S)(x,q)$, et il vient alors $\nabla^2_{q,x} S = \partial_x P = (\partial_p F)^{-1}$: donc $\nabla^2_{q,x} S$ et inversible et on peut définir $G(p,q) = \nabla_x S(F(p,q),q)$ de sorte que la transformation $(p,q) \to (x,y)$ soit symplectique. □

5.1 Définition d'un système globalement intégrable

Lorsque les composantes de F sont en involution deux à deux, alors, d'après le théorème 19 et la remarque 5, les flots $\varphi_t^{[F_i]}$ des systèmes

$$\begin{aligned} \dot{p} &= -\nabla_q F_i(p,q), \\ \dot{q} &= \nabla_p F_i(p,q), \end{aligned}$$

commutent deux à deux, et si $H_v(p,q) = v^T F(p,q)$ pour $v \in \mathbb{R}^d$, le flot $\varphi_t^{[H_v]}$ du système

$$\dot{p} = -\nabla_q H_v(p,q), \tag{5.1.1}$$
$$\dot{q} = \nabla_p H_v(p,q), \tag{5.1.2}$$

s'écrit

$$\varphi_t^{[H_v]} = \varphi_t^{[v_1 F_1]} \circ \cdots \circ \varphi_t^{[v_d F_d]} = \varphi_{v_1 t}^{[F_1]} \circ \cdots \circ \varphi_{v_d t}^{[F_d]} := \varphi_{vt}^{[F]}.$$

On a ainsi une structure de groupe commutatif[1]. Afin d'assurer l'existence de ce flot pour tout $t \in \mathbb{R}$ et tout $v \in \mathbb{R}^d$, on se place donc dans les conditions de la définition suivante :

Définition 4 *Soit H une fonction définie d'un ouvert U de $\mathbb{R}^d \times \mathbb{R}^d$ à valeur dans $\mathbb{R}$. Le système d'Hamiltonien H est dit complètement intégrable si il existe une fonction F, de U dans $\mathbb{R}^d$, de classe C^2, telle que :*

1. *Les composantes de F soient en involution deux à deux, c'est-à-dire :*

$$\forall\, (p,q) \in U,\ (\partial_p F)(\partial_q F)^T = (\partial_q F)(\partial_p F)^T. \tag{5.1.3}$$

[1]Notons que cette écriture est une conséquence de la commutativité des flots $\varphi_t^{[F_i]}$, elle-même conséquence du fait que les crochets de Poisson $\{F_i, F_j\}$ sont nuls deux à deux (les F_i sont en involution). En outre, les F_i sont alors des invariants du système (5.1.1) car $\frac{d}{dt} F_i \circ \varphi_t^{[H_v]} = \{F_i, H_v\} \circ \varphi_t^{[H_v]} = \sum_j v_j \{F_i, F_j\} \circ \varphi_t^{[H_v]} = 0$.

2. *La matrice $\partial_{(p,q)}F$ soit de rang d pour tout (p,q) de U.*
3. *Le flot $\varphi_v^{[F]}$ soit défini sur U pour tout $v \in \mathbb{R}^d$ et :*

$$\forall\,(p,q) \in U, \forall\, v \in \mathbb{R}^d,\ \varphi_v^{[F]}((p,q)) \in U. \tag{5.1.4}$$

Soit maintenant, pour $x \in F(U)$, M_x la surface de niveau x, c'est-à-dire :

$$M_x = F^{-1}(\{x\}) = \{(p,q) \in \mathbb{R}^d \times \mathbb{R}^d\ ;\ F(p,q) = x\}. \tag{5.1.5}$$

M_x est une sous-variété différentielle de $\mathbb{R}^{2d}$ de dimension d. Si pour $x \in F(U)$, la surface de niveau M_x est connexe, alors, sous les hypothèses de la définition 4, pour tout $(p_0,q_0) \in M_x$, l'application

$$\begin{array}{rcl} \sigma_{(p_0,q_0)}\ : \mathbb{R}^d & \rightarrow & M_x \\ v & \mapsto & \varphi_v^{[F]}(p_0,q_0) \end{array}$$

est surjective[2]. C'est une paramétrisation de M_x. Par la suite, on utilise le feuilletage

$$U = \bigcup_{x \in F(U)} M_x.$$

Théorème 14 *Soit F une fonction satisfaisant les hypothèses de la définition 4. Supposons que M_x soit connexe pour tout $x \in F(U)$. Alors il existe un ouvert $B \subset F(U)$ et une application symplectique et surjective e*

$$\begin{array}{rcl} e\ : B \times \mathbb{R}^d & \rightarrow & \bigcup_{x \in B} M_x \\ (x,y) & \mapsto & (p,q) \in M_x \end{array} \tag{5.1.6, 5.1.7}$$

telle que

$$\forall\, v \in \mathbb{R}^d, \forall\,(x,y) \in B \times \mathbb{R}^d,\ \varphi_v^{[F]}(e(x,y)) = e(x, y+v). \tag{5.1.8}$$

Preuve : Soient $(p_0,q_0) \in U$, $(x_0,y_0) = l(p_0,q_0)$ et B et W des voisinages de x_0 et y_0 respectivement tels que $l^{-1}(x,y)$ existe pour tout $x \in B$ et tout $y \in W$, où l désigne la transformation symplectique définie *localement* dans le lemme 7. On suppose désormais que $y_0 = 0$. Remarquons tout d'abord, que pour $\|v\|$ suffisament petit, le diagramme suivant commute :

$$\begin{array}{ccc} (p,q) \in M_x & \longrightarrow & \varphi_v^{[F]}(p,q) \in M_x \\ \uparrow l^{-1} & & \uparrow l^{-1} \\ (x,y) \in B \times W & \longrightarrow & (x, y+v) \end{array}$$

Sur $B \times W$, on peut donc définir l'application e comme l^{-1}. Mais pour étendre e à tous les $y \in \mathbb{R}^d$, il est nécessaire de la définir de la façon suivante :

$$\forall\, x \in B, \forall\, y \in \mathbb{R}^d,\ e(x,y) = \varphi_y^{[F]}(l^{-1}(x,0)). \tag{5.1.9}$$

On vérifie alors que :

[2]Pour tout point (p,q) de M_x et tout v de $\mathbb{R}^d$, $\sigma'_{(p,q)}(v) = J^{-1}(\partial_{(p,q)}F)^T$. Or $(\partial_{(p,q)}F)$ est de rang d donc $\sigma_{(p,q)}$ est un difféomorphisme local de $\mathbb{R}^d$ dans M_x. Maintenant si (p,q) est un point de M_x, il existe un chemin compact dans M_x reliant (p_0,q_0) à (p,q). $\sigma_{(p,q)}$ est un difféomorphime pour $\|v\| \le \delta$ uniformément en (p,q) le long de ce chemin. On peut donc découper ce chemin en un nombre fini de segments où σ est un difféomorphisme, chaque extrémité étant image de l'origine par l'application $\varphi_{v^i}^{[F]}$ pour un certain $\|v^i\| \le \delta$. En notant $v = \sum_i v^i$, on a alors $(p,q) = \sigma_{(p_0,q_0)}(v)$.

1. e **est surjective :** c'est une conséquence de la surjectivité de $\varphi_v^{[F]}(p,q) = \sigma_{(p,q)}(v)$ pour $(p,q) \in M_x$ connexe.
2. e **est symplectique :** au voisinage d'un point $y_1 \in \mathbb{R}^d$, on peut réécrire $e(x,y)$ pour $(y-y_1) \in W$ comme :

$$\begin{aligned} e(x,y) &= (\varphi_{y_1}^{[F]} \circ \varphi_{y-y_1}^{[F]} \circ l^{-1})(x,0), \\ &= (\varphi_{y_1}^{[F]} \circ l^{-1})(x, y-y_1). \end{aligned}$$

   Donc e s'écrit localement comme la composée de transformations symplectiques.

□

5.2 Variables actions-angles

Lorsque les surfaces de niveau sont compactes, on peut alors montrer le théorème d'Arnold-Liouville (1963), qui inscrit la dynamique de tout système hamiltonien intégrable de surfaces de niveau compactes sur un tore $\mathbb{T}^d$ de dimension d, paramétré par d angles modulo 2π, i.e. $\mathbb{T}^d = \mathbb{R}^d/2\pi\mathbb{Z}^d$. Lorsque un nombre $k < d$ de surfaces de niveau seulement sont compactes, alors la démontration donnée dans cette section montre que les trajectoires s'inscrivent sur un produit cartésien de tores et de droites réelles $\mathbb{T}^k \times \mathbb{R}^{d-k}$.

Théorème 15 (Arnold-Liouville) *Soit F une fonction satisfaisant les hypothèses de la définition 4. Supposons que M_x soit connexe et compacte pour tout $x \in F(U)$. Alors il existe un ouvert $B \subset F(U)$ et un ouvert D de $\mathbb{R}^d$ contenant 0 tels que :*

1. *Pour tout x de B, M_x est un tore de dimension d invariant par le flot $\varphi_{vt}^{[F]}$.*
2. *Il existe un difféomorphisme symplectique Ψ :*

$$\Psi \ : D \times \mathbb{T}^d \longrightarrow \bigcup_{x \in B} M_x \subset \mathbb{R}^d \times \mathbb{R}^d, \tag{5.2.1}$$

$$(a,\theta) \mapsto (p,q), \tag{5.2.2}$$

   tel que

$$\forall\, (a,\theta) \in D \times \mathbb{T}^d,\ (F \circ \Psi)(a,\theta) = f(a) \tag{5.2.3}$$

   où f est une fonction de D dans $\mathbb{R}^d$.

Preuve : Sous les hypothèses du théorème précédent, et pour $x \in B$, soit

$$\Gamma_x = e^{-1}(\{e(x,0)\} = \{y \in \mathbb{R}^d \ ; e(x,y) = e(x,0)\}. \tag{5.2.4}$$

Cet ensemble est un sous-groupe discret de $\mathbb{R}^d$. En effet :

1. Γ_x **est un groupe :** si y et y^* sont deux éléments de Γ_x, alors

$$e(x, y-y^*) = \varphi_{y-y^*}^{[F]}(e(x,0)) = \varphi_{y-y^*}^{[F]}(e(x,y^*)) = e(x, y^*+y-y^*) = e(x,y) = e(x,0),$$

   donc $y - y^* \in \Gamma_x$.

2. **Γ_x est discret :** pour y dans un voisinage d'un $y^* \in \Gamma_x$, e est un difféomorphisme (local), en particulier injectif, de sorte que $e(x,y) \neq e(x,y^*) = e(x,0)$, i.e. $y \notin \Gamma_x$. Donc y^* est un point isolé.

L'application $e(x,\cdot)$ est par contruction surjective et sa restriction

$$e(x,\cdot)\ : \mathbb{R}^d/\Gamma_x \ \longrightarrow \ M_x \tag{5.2.5}$$

est donc un difféomorphisme. La surface de niveau M_x étant supposée compacte, Γ_x est alors nécessairement une grille de dimension d, engendrée par d vecteurs $g_1(x), \ldots, g_d(x)$[3] :

$$\Gamma_x = \{\sum_{i=1}^{d} m_i g_i(x)\ ; m \in \mathbb{Z}^d\}. \tag{5.2.6}$$

Maintenant, on a pour tout $i = 1, \ldots, d$:

$$\forall x \in B,\ e(x, g_i(x)) = e(x,0)$$

soit en composant à gauche par l'application l définie sur $B \times W$, voisinage de $(x_0, 0)$:

$$\forall x \in B,\ l(e(x, g_i(x))) = l(e(x,0)) = (x,0).$$

En dérivant par rapport à x, et en notant $A = l'(e(x, g_i(x))) \circ e'(x, g_i(x))$, on obtient alors :

$$A \begin{bmatrix} I_{\mathbb{R}^d} \\ g_i'(x) \end{bmatrix} = \begin{bmatrix} I_{\mathbb{R}^d} \\ 0 \end{bmatrix},$$

de sorte que :

$$[I_{\mathbb{R}^d}\ g_i'(x)] A^T J A \begin{bmatrix} I_{\mathbb{R}^d} \\ g_i'(x) \end{bmatrix} = [I_{\mathbb{R}^d}\ 0] J \begin{bmatrix} I_{\mathbb{R}^d} \\ 0 \end{bmatrix} = 0.$$

L'application $l \circ e$ étant symplectique, $A^T J A = J$, d'où finalement :

$$(g_i'(x))^T = g_i'(x). \tag{5.2.7}$$

Le lemme d'intégrabilité permet alors d'affirmer que $g_i(x) = \nabla_x U_i(x)$ où l'on peut supposer que $U_i(x_0) = 0$ et le difféomorphisme $e(x,\cdot)$ induit par e s'écrit :

$$\begin{array}{rcl} e(x,\cdot)\ : \mathbb{T}^d & \longrightarrow & M_x \\ \theta & \longmapsto & e(x, \frac{1}{2\pi} \sum_i \theta_i \nabla_x U_i(x)). \end{array}$$

Il reste désormais à étendre $e(x,\cdot)$ en une application symplectique. Pour cela, on considère la fonction génératrice :

$$S(x,\theta) = \frac{1}{2\pi} \sum_i \theta_i U_i(x)). \tag{5.2.8}$$

[3]Les fonctions g_i peuvent être définies dans un voisinage de x_0 comme des fonctions C^1. En effet, soit $\tilde{l}$ l'inverse de e au voisinage de $(x_0, g_i(x_0))$. En posant $(\alpha(x), \beta(x)) = (\tilde{l} \circ e(x,0))$, on définit localement deux fonctions C^1. De $(F \circ e)(x,y) = x$, et de $e(\alpha(x), \beta(x)) = e(x,0)$ on tire alors $\alpha(x) \equiv x$ de sorte que l'on peut définir $g_i(x) = \beta(x)$. On suppose par la suite que le voisinage de x_0 où les g_i sont bien définies et C^1 se confond avec B (quitte à restreindre B).

Puisque $\nabla^2_{x,\theta} S(x,\theta) = \frac{1}{2\pi}[g_1(x), \ldots, g_d(x)]$ est inversible, les équations

$$a = \nabla_\theta S(x,\theta) = \frac{1}{2\pi}[U_1(x), \ldots, U_d(x)]^T \text{ et } y = \nabla_x S(x,\theta) = \frac{1}{2\pi}\sum_i \theta_i g_i(x)$$

définissent une transformation symplectique ϑ (d'après le lemme 6) :

$$\begin{array}{rcl} \vartheta\ : D \times \mathbb{R}^d & \longrightarrow & B \times \mathbb{R}^d \\ (a,\theta) & \longmapsto & (x,y) = \left(f(a), \dfrac{1}{2\pi}\displaystyle\sum_i \theta_i g_i(f(a)) \right) \end{array}$$

La composée

$$\Psi = e \circ \vartheta\ : D \times \mathbb{R}^d \ \longrightarrow \ \bigcup_{x\in B} M_x$$

est par construction surjective et symplectique, sa restriction à $D \times \mathbb{T}^d$ est injective, et $(F \circ \Psi)(a,\theta) = (F \circ e)\left(f(a), \frac{1}{2\pi}\sum_i \theta_i g_i(f(a))\right) = f(a)$. $\square$

Chapitre 6

Schémas symplectiques et théorie KAM

6.1 Ergodicité des systèmes complètements intégrables

Les systèmes hamiltoniens complètement intégrables se résolvent par "quadrature" au sens où une fois connus les invariants, il est possible de construire à l'aide de fonctions génératrices un système de variables actions-angles $(a, \theta) \in \mathbb{T}^d \times \mathbb{R}^d$, qui permettent de calculer à la main la solution du système. Dans ces variables, le système devient

$$\left\{ \begin{array}{rcl} \dot{a} & = & 0, \\ \dot{\theta} & = & \omega(a), \end{array} \right. \tag{6.1.1}$$

où $\omega(a) = \nabla_a K(a)$ où $K(a)$ est le Hamiltonien. Notons si une des composantes de $\omega(a)$ s'annule, le système est dit "super intégrable", c'est-à-dire qu'il possède un invariant de plus. C'est le cas du système de Kepler en dimension 2 avec le potentiel à champ central $V(r) = -1/r$.

Le système dynamique (6.1.1) décrit les géodésiques du tore $\mathbb{T}^d$ de vecteurs tangent initial $\omega(a)$. Intuitivement, il est clair que plusieurs situations peuvent se présenter selon que les composantes du vecteur $\omega := \omega(a)$ sont "rationnellement dépendantes" ou pas. Considérons en effet le cas $d = 2$, et supposons qu'il existe des entiers k_1 et k_2 tels que $k_1/k_2 = \omega_1/\omega_2$. Alors on a $k_1/\omega_1 = k_2/\omega_2$ ce qui implique que pour le temps $t = 2\pi(k_1/\omega_1)$ alors on a $t\omega_1 = 2k_1\pi$ et $t\omega_2 = 2k_2\pi$, c'est-à-dire dans le tore $\mathbb{T}^2$, $t\omega = 0$, et la trajectoire $t \to (a_0, \theta_0 + t\omega)$ est périodique. En revanche, quand ω_1 et ω_2 sont rationnellement indépendant, l'intuition est que la trajectoire "remplit" le tore, au sens où tout point du tore est atteint (éventuellement en temps très long) par cette trajectoire.

En dimension d, cette situation se produit (dans un sens plus faible) lorsque les fréquences sont rationnellement indépendantes, ou "non résonantes", c'est-à-dire satisfont la propriété

$$\forall\, (k_1, \ldots, k_d) \in \mathbb{Z}^d \quad k_1\omega_1 + \cdots + k_d\omega_d = 0 \quad \Rightarrow \quad k_1 = \cdots = k_d = 0. \tag{6.1.2}$$

Dans la suite, on note $k \cdot \omega = k_1\omega_1 + \cdots + k_d\omega_d$. Sous cette condition, alors on montre que les trajectoires de (6.1.1) sont denses dans le tore :

Théorème 16 (Arnold) *Soit f une fonction continue sur le tore $\mathbb{T}^d$ (en particulier elle est périodique). On définit la moyenne spatiale de f sur $\mathbb{T}^d$ par l'équation*

$$\langle f \rangle := \frac{1}{(2\pi)^d} \int_0^{2\pi} \cdots \int_0^{2\pi} f(\theta) d\theta_1 \cdots d\theta_d. \tag{6.1.3}$$

Soit $\theta_0 \in \mathbb{T}^d$. Alors sous l'hypothèse (6.1.2) que les fréquences sont non résonnantes, on a

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T f(\theta_0 + t\omega)dt = \langle f\rangle.$$

Ce résultat dit que la moyenne temporelle de f le long de la solution de (6.1.1) (pour un a_0 fixé) est égale à la moyenne spatiale de f sur le tore. Cette propriété caractérise le fait que le système (6.1.1) est "ergodique".
Preuve : On montre tout d'abord le théorème pour f un polynôme trigonométrique de la forme[1]

$$f = \sum_{k\in\mathbb{Z}^d,\, |k|\le N} f_k e^{ik\cdot\theta}.$$

On calcule alors que

$$\begin{aligned}\frac{1}{T}\int_0^T f(\theta_0 + t\omega)dt &= \frac{1}{T}\sum_{|k|\le N} f_k \int_0^T e^{itk\cdot\omega}dt\\ &= f_0 + \frac{1}{T}\sum_{|k|\le N,\, k\neq 0} f_k \frac{e^{iTk\cdot\omega}-1}{ik\cdot\omega},\end{aligned}$$

où on a utilisé l'hypothèse (6.1.2). On en déduit donc que

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T f(\theta_0+t\omega)dt = f_0,$$

or il est clair que

$$\langle f\rangle = \sum_{|k|\le N} f_k \frac{1}{(2\pi)^d}\int_0^{2\pi}\cdots\int_0^{2\pi} e^{ik\cdot\theta}d\theta_1\cdots d\theta_d = f_0,$$

ce qui prouve le résultat pour les polynômes trigonométriques.

Soit maintenant f une fonction continue sur $\mathbb{T}^d$ et soit $\varepsilon > 0$. Il existe un polynôme trigonométrique P tel que $\|f-P\|_{L^\infty(\mathbb{T}^d)} < \frac{1}{2}\varepsilon$. On pose $P_1 = P - \frac{1}{2}\varepsilon$ et $P_2 = P + \frac{1}{2}\varepsilon$. On a $P_1 < f < P_2$ et $\langle P_2 - P_1\rangle = \varepsilon$. Pour tout $T > 0$ on a

$$\frac{1}{T}\int_0^T P_1(\theta_0+t\omega)dt < \frac{1}{T}\int_0^T f(\theta_0+t\omega)dt < \frac{1}{T}\int_0^T P_2(\theta_0+t\omega)dt.$$

Or il existe $T(\varepsilon)$ tel que pour $T > T(\varepsilon)$ on a

$$\left|\frac{1}{T}\int_0^T P_i(\theta_0+t\omega)dt - \langle P_i\rangle\right| < \varepsilon \quad \text{pour} \quad i = 1, 2.$$

On en déduit donc que pour $T > T(\varepsilon)$ on a

$$\langle P_1\rangle - \varepsilon < \frac{1}{T}\int_0^T f(\theta_0+t\omega)dt < \langle P_2\rangle + \varepsilon$$

Or on a $\langle P_1\rangle < \langle f\rangle < \langle P_2\rangle$, ce qui s'écrit encore

$$-\langle P_2\rangle < -\langle f\rangle < -\langle P_1\rangle.$$

[1]Pour un multi-indice $k \in \mathbb{Z}^d$, on note dans la suite $|k| = |k_1| + \cdot + |k_d|$ et $k\cdot\theta = k_1\theta_1 + \ldots + k_d\theta_d$.

En sommant les deux précédentes inégalités, on trouve

$$\langle P_1 - P_2\rangle - \varepsilon < \frac{1}{T}\int_0^T f(\theta_0 + t\omega)dt - \langle f\rangle < \langle P_2 - P_1\rangle + \varepsilon.$$

Or $\langle P_2 - P_1\rangle = \varepsilon$. On en déduit donc que pour $T > T(\varepsilon)$ on a

$$\left|\frac{1}{T}\int_0^T f(\theta_0 + t\omega)dt - \langle f\rangle\right| \le 2\varepsilon,$$

ce qui montre le résultat. □

Corollaire 1 *Soit un système hamiltonien complètement intégrable, et soient F et M_x satisfaisant les hypothèses du théorème d'Arnold-Liouville (en particulier la compacité de la surface de niveau M_x). Alors pour toute fonction continue A sur l'espace des phase et pour tout $(p_0, q_0) \in M_x$ on a*

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T A(p(t), q(t))dt = \frac{\int_{M_x} A(p,q)d\sigma(p,q)}{\int_{M_x} d\sigma(p,q)},$$

où $d\sigma(p,q)$ est l'élément de volume "normalisé"[2] sur M_x induit par l'espace euclidien $\mathbb{R}^{2d}$.

Cette propriété d'ergodicité des systèmes complètement intégrables pour des fréquences non résonnantes est un premier pas dans l'étude de systèmes dynamiques ergodiques et constitue un des rares exemples où on sait caractériser l'ergodicité. Notons que l'étude de l'ergodicité de systèmes dynamiques issus de la mécanique a des applications pratiques très importantes en dynamique moléculaire et en astronomie.

6.2 Theorie classique des perturbations hamiltoniennes

Dans le but d'étudier d'autres systèmes plus généraux que les systèmes complètement intégrables, les scientifiques du 19$^{\text{ème}}$ siècle (Poincaré, Lindstedt) ont d'abord considéré des "perturbations" de systèmes complètement intégrables.

Il apparait ces systèmes sont stables par perturbations du hamiltonien : c'est tout l'objet de la théorie KAM développée par Birkhoff, Siegel, Kolmogorov/Arnold/Moser et Nekhoroshev au cours du 20$^{\text{ème}}$ siècle.

Considérons un système complètement intégrable tel que pour des variables actions-angles (a, θ) dans un domaine $D \times \mathbb{T}^d$ $(D \subset \mathbb{R}^d)$, le hamiltonien s'écrive $H_0(a)$. Considérons une petite perturbation de ce hamiltonien

$$H_\varepsilon(a, \theta) = H_0(a) + \varepsilon H_1(a, \theta). \tag{6.2.1}$$

La problématique de la théorie KAM peut se poser ainsi :

Peut-on trouver un changement de variable symplectique $(a, \theta) \to (b, \varphi)$ proche de l'identité tel que dans ces nouvelles variables, on ait $H_\varepsilon(a, \theta) = K(b)$, c'est-à-dire que le système est de nouveau complètement intégrable.

Notons qu'il est important que cette perturbation ait lieu sur le seul hamiltonien, c'est-à-dire que la perturbation ne modifie pas la structure symplectique sous jacente.

[2] Dans un sens qu'on ne précise pas ici

Les résultat de l'analyse rétrograde pour une méthode symplectique montrent que la solution numérique associée à un système hamiltonien est exponentiellement proche d'une système hamiltonien modifié dont le hamiltonien s'écrit

$$H(p,q) + h^r H_h(p,q),$$

où h est le pas de temps, r l'ordre de la méthode et H_h un polynôme en h. Le but de ce chapitre est donc de combiner les résultats de la théorie KAM à ceux de l'analyse rétrograde pour obtenir des renseignements sur le comportement de la solution numérique en temps long.

6.2.1 Equations en séries formelles

On suppose maintenant que les fonctions H_0 et H_1 de (6.2.1) sont analytiques sur le domaine $D \times \mathbb{T}^d$. Les résultats qui suivent ont été adaptés par Moser pour des fonctions moins régulières, mais toujours continuement différentiables à des ordres élevés.

Afin d'obtenir un changement de variable $(a,\theta) \to (b,\varphi)$ qui soit symplectique, on le cherche tout d'abord à l'aide d'une fonction génératrice $S(b,\theta)$ "proche de l'identité" au sens où

$$S(b,\theta) = b\cdot\theta + \varepsilon S_1(b,\theta) + \varepsilon^2 S_2(b,\theta) + \cdots$$

et donc

$$a = b + \varepsilon\nabla_\theta S_1(b,\theta) + \varepsilon^2\nabla_\theta S_2(b,\theta) + \cdots \quad \text{et} \quad \varphi = \theta + \varepsilon\nabla_b S_1(b,\theta) + \varepsilon^2\nabla_b S_2(b,\theta) + \cdots$$

Considérons une série formelle

$$S[\varepsilon](b,\theta) = b\cdot\theta + \sum_{i\geq 1}\varepsilon^i S_i(b,\theta)$$

donc les coefficients $S_i(b,\theta)$ sont analytiques sur un domaine $D\times\mathbb{T}^d$. On peut définir la série formelle

$$a[\varepsilon] = b + \sum_{i\geq 1}\varepsilon^i\nabla_\theta S_i(b,\theta).$$

Considérons un point b fixé. A l'aide de développements de Taylor, on peut écrire au sens des séries formelles que

$$H_0(a) + \varepsilon H_1(a,\theta) = H_\varepsilon(a,\theta) = \sum_{k\geq 0}\frac{1}{k!}\nabla_a^k H_\varepsilon(b,\theta)\cdot(a-b,\ldots,a-b),$$

où $(a-b)$ apparaît k fois dans l'argument de $\nabla_a^k H_\varepsilon(b,\theta)$. Rappelons que pour i un entier et pour une fonction $f(a,\theta)$ sur $D\times\mathbb{T}^d$, on définit la différentielle d'ordre n par rapport à a comme l'application $\nabla_a^n f$ qui aux n champs de vecteurs $X_1,\ldots,X_n$ de composantes $X_i^{(l)}$ pour $l = 1,\ldots,d$ et $i = 1,\ldots,n$ fait correspondre la fonction

$$(\nabla_a^n f)(X_1,\ldots,X_n) := \sum_{1\leq l_j\leq d}\frac{\partial^n f}{\partial a_{l_1}\ldots\partial a_{l_n}}X_1^{(l_1)}\cdots X_n^{(l_n)}.$$

On peut donc définir la série formelle

$$\mathcal{H}[\varepsilon](b,\theta) := H_\varepsilon(a[\varepsilon],\theta) = \sum_{k\in\mathbb{N}}\frac{1}{k!}\nabla_a^k H_\varepsilon(b,\theta)\cdot(a[\varepsilon]-b,\ldots,a[\varepsilon]-b) =: \sum_{i\geq 0}\varepsilon^i\mathcal{H}_i(b,\theta)$$

en utilisant le développement de Taylor précédent et le développement de la série formelle $a[\varepsilon]$, puis en identifiant directement les puissances de ε. On calcule directement que (rappelons que $\omega(a) = \nabla_a H_0(a)$),

$$\begin{aligned}
\mathcal{H}_0(b,\theta) &= H_0(b) \\
\mathcal{H}_1(b,\theta) &= \omega(b)\cdot\nabla_\theta S_1(b,\theta) + H_1(b,\theta) \\
\mathcal{H}_2(b,\theta) &= \omega(b)\cdot\nabla_\theta S_2(b,\theta) + \nabla_a H_1(b,\theta)\cdot\nabla_\theta S_1(b,\theta) + \tfrac{1}{2}\nabla_a^2 H_0(b)\left(\nabla_\theta S_1(b,\theta), \nabla_\theta S_1(b,\theta)\right).
\end{aligned}$$

Plus généralement, on a pour tout $j \geq 1$

$$\mathcal{H}_j(b,\theta) = \omega(b)\cdot\nabla_\theta S_j(b,\theta) + K_j(b,\theta), \tag{6.2.2}$$

où $K_1 = H_1$ et pour tout $j \geq 2$,

$$K_j = \sum_{i=2}^{j} \sum_{k_1+\ldots+k_i=j} \frac{1}{i!}\nabla_a^i H_0\left(\nabla_\theta S_{k_1},\ldots,\nabla_\theta S_{k_i}\right) + \sum_{i=1}^{j-1} \sum_{k_1+\ldots+k_i=j-1} \frac{1}{i!}\nabla_a^i H_1\left(\nabla_\theta S_{k_1},\ldots,\nabla_\theta S_{k_i}\right), \tag{6.2.3}$$

où la sommation se fait sur des k_i entier plus grand ou égal à 1 (la même convention sera valable dans la suite).

Le problème en série formelle associé au problème initial s'énonce donc de la façon suivante :

Peut-on trouver une série formelle $S[\varepsilon](b,\theta)$ à coefficients analytiques sur $D\times\mathbb{T}^d$ telle que les coefficients de la série formelle $\mathcal{H}[\varepsilon](b,\theta)$ ne dépendent que de b

Pour $f(a,\theta)$ une fonction sur $D\times\mathbb{T}^d$, on définit la moyenne de f sur le tore comme

$$\langle f\rangle(a) = \frac{1}{(2\pi)^d}\int_{\mathbb{T}^d} f(a,\theta)d\theta$$

où l'intégrale sur le tore est celle de la formule (6.1.3). On peut donc reformuler le problème en écrivant qu'on cherche des fonctions $S_k(b,\theta)$ telles que

$$\mathcal{H}[\varepsilon](b,\theta) = \langle\mathcal{H}[\varepsilon]\rangle(b).$$

En utilisant l'équation (6.2.2) cette équation en série formelle est donc équivalente à la collection d'équations :

$$\forall\, j \geq 1, \quad \omega(b)\cdot\nabla_\theta S_j(b,\theta) + K_j(b,\theta) = \langle K_j\rangle(b), \tag{6.2.4}$$

puisque $\langle\omega(b)\cdot\nabla_\theta S_j(b,\theta)\rangle = 0$. Ce type d'équation est à la base de la théorie des perturbations hamiltoniennes.

6.2.2 Résolution des équations

Pour résoudre, au moins formellement les équations (6.2.4), on se place dans l'espace des fonctions analytiques sur $D\times\mathbb{T}^d$ et on cherche des développements formels des fonctions S_j en séries de Fourier. Considérons l'équation (6.2.4) pour $j=1$. Elle s'écrit

$$\omega(b)\cdot\nabla_\theta S_1(b,\theta) + H_1(b,\theta) = \langle H_1\rangle(b).$$

Supposons que H_1 est analytique. On peut écrire

$$H_1(b,\theta) = \sum_{k\in\mathbb{Z}^d} h_k(b)e^{ik\cdot\theta}$$

et les coefficients $h^k(b)$ décroissent exponentiellement en fonction de k (voir ci-dessous pour un énoncé précis).

Si on cherche $S_1(b,\theta)$ sous la forme

$$S_1(b,\theta) = \sum_{k\in\mathbb{Z}^d} s_k(b)e^{ik\cdot\theta}$$

on est conduit aux équations pour les coefficients de Fourier :

$$\forall\, k \in \mathbb{Z}^d, \quad k \neq 0, \quad (i\omega(b)\cdot k)s_k(b) + h_k(b) = 0.$$

On peut les résoudre sous l'hypothèse que les fréquences sont non résonnantes par l'équation

$$\forall\, k \in \mathbb{Z}^d, \quad k \neq 0, \quad s_k(b) = -\frac{h_k(b)}{ik\cdot\omega(b)}.$$

On voit que des problèmes peuvent survenir si le produit $k\cdot\omega(b)$ devient trop petit (problème des "petits dénominateurs"). Une des possibilités pour éviter ces problèmes est de faire l'hypothèse que $\omega = \omega(b)$ satisfait la *condition diophantienne de Siegel* suivante : il existe $\gamma > 0$ et $\nu > d-1$ tels que

$$\forall\, k \in \mathbb{Z}^d, \quad k \neq 0, \quad |k\cdot\omega| \geq \gamma|k|^{-\nu}. \tag{6.2.5}$$

Cette équation est encore appelée *condition de non résonnance forte* car elle implique en particulier le théorème d'Arnold pour le calcul de moyennes.

Ainsi, si H_1 est analytique, alors $h_k(b) \leq Ce^{-\rho|k|}$ pour tout $k \neq 0$, et la condition de Siegel implique que les coefficients $s_k(b)$ sont encore exponentiellement décroissants et donc qu'on peut trouver S_1 solution analytique de l'équation.

La condition de Siegel est vraie "presque partout" au sens suivant : presque toutes les fréquences satisfont (6.2.5) pour un $\gamma > 0$. C'est une conséquence du lemme suivant :

Lemme 8 *Soit $\Omega \subset \mathbb{R}^d$ un domaine borné, et soit deux réels $\nu > d-1$ et $\gamma > 0$. Alors la mesure de Lebesgue de l'ensemble*

$$\{\omega \in \Omega\ ;\ \omega \textit{ ne satisfait pas } (6.2.5)\}$$

est inférieure à $C(\Omega,d,\nu)\gamma$ où la constante $C(\Omega,d,\nu)$ ne dépend que du domaine Ω et de la dimension d et de ν.

Preuve : Considérons un k fixé et un $\omega \in \Omega$. La norme $|k|$ est la norme 1 de k sur $\mathbb{R}^d$ qui est donc équivalente à la norme euclidienne. Il est clair qu'il existe $\omega_0 \in \mathbb{R}^d$ et un α réel tel que $k\cdot\omega_0 = 0$ et

$$\omega = \omega_0 + \alpha\frac{k}{|k|}. \tag{6.2.6}$$

Soit ω ne satisfaisant pas la condition (6.2.5), alors il existe k tel que $|k\cdot\omega| \leq \gamma|k|^{-\nu}$. Pour ce ω et ce k, on a donc $|\alpha| \leq c\gamma|k|^{-\nu-1}$ où c ne dépend que de d.

Ainsi, l'ensemble des ω ne satisfaisant pas (6.2.5) et appartenant à Ω est inclu dans l'ensemble Z des $\omega := \omega_0 + \alpha k/|k| \in \Omega$ pour $k \in \mathbb{Z}^d$, tels que ω_0 dans l'orthogonal de k, et α de valeur absolue plus petite que $c\gamma|k|^{-\nu-1}$.

Or puisque Ω est borné, on a que $\omega_0 = \omega - \alpha k/|k|$ est borné pour la norme euclidienne sur $\mathbb{R}^d$ indépendemment de k. Donc pour tout k, la mesure de l'ensemble des ω_0 possible est indépendante du k choisi. Ainsi, la mesure de Lesbesgue de Z est plus petite que

$$\gamma C \sum_{k\in\mathbb{Z}^d, k\neq 0} |k|^{-\nu-1}$$

où C ne dépend que de Ω et de la dimension d.

Enfin, on montre que

$$\sum_{k\in\mathbb{Z}^d, k\neq 0} |k|^{-\nu-1} < \infty \quad \text{si} \quad \nu > d-1.$$

En effet, on a

$$\sum_{k\in\mathbb{Z}^d, k\neq 0} |k|^{-\nu-1} = 2^d \sum_{k\in\mathbb{N}^d,\, k\neq 0} \frac{1}{|k|^{\nu+1}} = 2^d \sum_{n\geq 1}\sum_{k\in S_n} \frac{1}{n^{\nu+1}} = 2^d \sum_{n\geq 1} \mathrm{Card}(S_n)\frac{1}{n^{\nu+1}}$$

où S_n est l'ensemble :

$$S_n = \{k\in\mathbb{N}^d\,;\, k_1+\ldots+k_d = n\},$$

dont le cardinal est

$$\mathrm{Card}(S_n) = \binom{n+d-1}{n} = \frac{(n+d-1)\cdots(n+1)}{(d-1)!} \leq Cn^{d-1}.$$

Il est donc clair que la série de terme général $\mathrm{Card}(S_n)n^{-\nu-1}$ converge si $\nu > d-1$. □

Dans toute la suite, ν désignera un entier plus grand que d.

Avant de montrer un résultat précis concernant la résolution des équations (6.2.4), nous donnons quelques définitions et notations. Considérons une fonction F réelle analytique sur $[0,2\pi]$. Par définition, il existe M et R tels que pour tout $k\in\mathbb{N}$, on ait

$$\forall\,\theta\in\mathbb{T}^d, \quad |F^{(k)}(\theta)| \leq Mk!R^k.$$

Ceci implique de manière équivalente que F est développable en série entière au voisinage de tout point. Comme $[0,2\pi]$ est compact, on peut donc trouver un réel ρ_0 tel que F est développable en série entière au voisinage de tout point de $[0,2\pi]$ avec un rayon de convergence égal à ρ_0. On peut donc définir la fonction d'une variable complexe $F(z)$ dans l'ouvert $\{z\in[0,2\pi]+i\mathbb{R}\,;\,|\Im z|<\rho_0\}$.

Considérons les coefficients de Fourier de F :

$$f_k = \frac{1}{2\pi}\int_0^{2\pi} F(\theta)e^{-ik\theta}d\theta.$$

Pour les estimer, on considère le contour Γ du plan complexe définit par

$$\Gamma = [0,2\pi]\cup(2\pi+i[-\rho,0])\cup([0,2\pi]-i\rho)\cup i[-\rho,0]$$

parcouru dans le sens des aiguilles d'une montre, où $0<\rho<\rho_0$. On a alors pour tout k,

$$\int_\Gamma F(z)e^{-ikz}dz = 0$$

car la fonction $z\to g(z) = F(z)e^{-ikz}$ est holomorphe. On en déduit que

$$\int_0^{2\pi} g(\theta)d\theta - \int_{-\rho}^0 g(2\pi+it)dt - \int_0^{2\pi} g(\theta-i\rho)d\theta + \int_{-\rho}^0 g(it)dt = 0.$$

Le premier terme de cette somme est le coefficient de Fourier f_k multiplié par 2π, et le troisième et le quatrième se compensent car F est périodique (on vérifie que F étendue au complexe est effectivement périodique !). On en déduit que

$$f_k = \frac{1}{2\pi}\int_0^{2\pi} F(\theta-i\rho)e^{-ik\theta-\rho k}d\theta$$

et donc

$$|f_k| \leq \max_{|\Im z| \leq \rho} |F(z)| e^{-\rho k}.$$

Ces résultats s'étendent immédiatement aux fonctions réelles analytiques sur le tore $\mathbb{T}^d$. Dans la suite, pour $\rho > 0$, on pose

$$U_\rho = \{\theta \in \mathbb{T}^d + i\mathbb{R}^d \,;\, \|\Im\theta\| < \rho\}$$

où $\|\cdot\|$ désigne la norme sup sur $\mathbb{R}^d$. Il est alors clair que si F est analytique sur le tore, il existe un ρ_0 tel que $F(\theta)$ désigne une fonction développable en série entière sur U_{ρ_0}.

Dans la suite, on note pour tout $\rho > 0$,

$$\|F\|_\rho = \sup_{\theta \in U_\rho} |F(\theta)|.$$

De la même manière, si F est réelle analytique sur le tore, il en est de même de $\nabla_\theta F$, avec le même rayon de convergence global ρ_0. On peut donc définir

$$\|\nabla_\theta F\|_\rho = \sum_{j=1}^{d} \left\| \frac{\partial F}{\partial \theta_j} \right\|_\rho .$$

Enfin, on peut définir les coefficients de Fourier

$$f_k = \frac{1}{(2\pi^d)} \int_{\mathbb{T}^d} F(\theta) e^{-ik\cdot\theta} d\theta$$

pour tout $k \in \mathbb{Z}^d$, et on montre de la même manière que précédemment que pour tout $\rho < \rho_0$, on a

$$\forall\, k \in \mathbb{Z}^d, \quad |f_k| \leq \|F\|_\rho \, e^{-\rho|k|}.$$

Le principe de la résolution des équations (6.2.4) est donné dans la proposition suivante :

Proposition 1 *Soit $\omega \in \mathbb{R}^d$ satisfaisant la condition diophantienne (6.2.5). Soit G une fonction réelle analytique sur un domaine U_ρ pour $\rho > 0$, et soit $\langle G \rangle$ la moyenne de G sur le tore. Alors l'équation*

$$\omega \cdot \nabla_\theta F + G = \langle G \rangle$$

possède une unique solution réelle analytique F sur U_ρ de moyenne nulle ($\langle F \rangle = 0$). De plus, pour tout $\delta < \min(\rho, 1)$, F est bornée sur $U_{\rho-\delta}$ par

$$\|F\|_{\rho-\delta} \leq \kappa_0 \delta^{-\alpha+1} \|G\|_\rho \quad \text{et} \quad \|\nabla_\theta F\|_{\rho-\delta} \leq \kappa_1 \delta^{-\alpha} \|G\|_\rho ,$$

où $\alpha = \nu + d + 1$, $\kappa_0 = \gamma^{-1} 8^d 2^{\nu+1} \nu!$ et $\kappa_1 = \gamma^{-1} 8^d 2^{\nu+2} (\nu+1)!$.

Preuve : Pour $\theta \in U_\rho$, on peut considérer la série de Fourier

$$G(\theta) - \langle G \rangle = \sum_{k \in \mathbb{Z}^d, k \neq 0} g_k e^{ik\cdot\theta}.$$

Cette série est convergente dans U_ρ car on a

$$|g_k| \leq M e^{-|k|\rho}$$

avec $M = \|G - \langle G\rangle\|_\rho \leq 2\|G\|_\rho$.

Cherchons F sous la forme

$$F(\theta) = \sum_{k\in\mathbb{Z}^d} f_k e^{ik\cdot\theta}, \tag{6.2.7}$$

avec $f_0 = \langle F\rangle = 0$. On trouve donc les équations :

$$\forall\, k \in \mathbb{Z}^d, \quad k \neq 0, \quad f_k = -\frac{g_k}{ik\cdot\omega}.$$

La condition sur les coefficients g_k implique que la série (6.2.7) définit bien une fonction F sur U_ρ dont la restriction à $\mathbb{T}^d$ est réelle analytique. De plus, on a

$$\|F\|_{\rho-\delta} \leq \sum_k |f_k| e^{|k|(\rho-\delta)} \leq \frac{M}{\gamma}\sum_k |k|^\nu e^{-|k|\delta}.$$

On utilise alors l'inégalité $x^\nu/\nu! \leq e^x$ pour $x = |k|\delta/2$, et on trouve

$$\sum_k |k|^\nu e^{-|k|\delta} = \sum_k |k|^\nu e^{-|k|\delta/2} e^{-|k|\delta/2} \leq 2^\nu \delta^{-\nu}\nu! \sum_k e^{-|k|\delta/2}.$$

Cette dernière somme s'écrit encore

$$\sum_k e^{-|k|\delta/2} = \sum_{k_1,\dots,k_d\in\mathbb{Z}} e^{-|k_1|\delta/2}\cdots e^{-|k_d|\delta/2} = \Big(1 + 2\sum_{j=1}^\infty e^{-j\delta/2}\Big)^d.$$

On calcule alors directement :

$$\Big(1 + 2\sum_{j=1}^\infty e^{-j\delta/2}\Big)^d = \Big(\frac{2}{1-e^{-\delta/2}} - 1\Big)^d = \Big(\frac{1+e^{-\delta/2}}{1-e^{-\delta/2}}\Big)^d.$$

Or si δ est assez petit ($\delta < 1$), on a

$$\frac{1+e^{-\delta/2}}{1-e^{-\delta/2}} \leq \frac{8}{\delta}.$$

Finalement, en groupant des estimations, on trouve

$$\|F\|_{\rho-\delta} \leq M\gamma^{-1}2^\nu\delta^{-\nu}\nu! 8^d\delta^{-d} \leq (\gamma^{-1}8^d 2^{\nu+1}\nu!)\delta^{-\nu-d}\|G\|_\rho,$$

ce qui montre la première estimation. Pour la deuxième, on calcule directement que

$$\|\nabla_\theta F\|_{\rho-\delta} \leq \sum_k |f_k|\cdot|k| e^{|k|(\rho-\delta)} \leq \frac{M}{\gamma}\sum_k |k|^{\nu+1} e^{-|k|\delta},$$

et on déduit le résultat en utilisant la borne précédente avec ν remplacé par $\nu+1$. $\square$

Ce résultat donne immédiatement l'existence de la fonction S_1. Dans toute la suite, on suppose que $\rho < 1$ afin de pouvoir toujours appliquer cette proposition. En utilisant l'analyticité de H_0 et H_1, on montre par récurrence l'existence des des fonctions S_k solutions de (6.2.4), pour un b^* satisfaisant la condition diophantienne. Plus précisemment, on a le résultat suivant :

Proposition 2 *Soit $b^* \in \mathbb{R}^d$ tel que H_0 soit développable en série entière dans un voisinage complexe $B_r(b^*)$ de b^*, où r est un nombre réel positif*[3]*. Supposons de plus que H_1 soit développable en série entière sur le domaine $B_r(b^*) \times U_\rho$, et que $\omega(b^*)$ satisfait la condition diophantienne* (6.2.5)*. Alors il existe des fonctions réelles analytiques S_j sur $\mathbb{T}^d$, dépendant de b^*, notées $S_j(b^*, \cdot)$, et telles que pour tout $j \geq 1$, on ait*

$$\|\nabla_\theta S_j(b^*, \cdot)\|_{\rho/2} \leq C_0(C_1 j^{2\alpha})^j,$$

où $C_0 = 3r/2$, $C_1 = 8(2^\alpha \kappa_1 M / r\rho^\alpha)^2$, où κ_1 et α sont donnés par la proposition précédente, et où

$$M = \max\left(\sup_{B_r(b^*)} |H_0|, \sup_{B_r(b^*) \times U_\rho} |H_1| \right). \tag{6.2.8}$$

Preuve : Rappelons que les S_j sont déterminés par récurrence par les équations (6.2.4), où les K_j sont donnés par (6.2.3). Puisque H_0 et H_1 sont développables en séries entières, il est clair en utilisant l'expression des K_j ainsi que la proposition précédente qu'il existe une suite S_j de fonctions réelles analytiques sur le tore.

Afin d'obtenir l'estimation de la proposition, fixons un indice J, et posons $\delta = \rho/(2J)$. Dans la suite, on pose pour tout $k \geq 1$ et pour j tel que $1 \leq j \leq J$,

$$\|K_k\|_{(j)} := \|K_k(b^*, \cdot)\|_{\rho - j\delta},$$

et une notation similaire pour $\nabla_\theta S_j(b^*, \cdot)$. Pour tout $j \geq 1$, $S_j(b^*, \cdot)$ est solution de l'équation

$$\omega(b^*) \cdot \nabla_\theta S_j(b^*, \cdot) + K_j(b^*, \theta) = \langle K_j \rangle (b^*),$$

et d'après la proposition précédente, on a donc pour tout j,

$$\|\nabla_\theta S_j\|_{(j)} \leq \kappa_1 \delta^{-\alpha} \|K_j\|_{(j-1)}. \tag{6.2.9}$$

La fonction H_0 étant analytique dans $B_r(b^*)$, on a par définition de M que pour tout $v_1, \ldots, v_d$ dans $\mathbb{C}^d$ et tout $i \geq 0$,

$$\left| \frac{1}{i!} \nabla_a^i H_0(b^*)(v_1, \ldots, v_d) \right| \leq \frac{M}{r^i} |v_1| \cdots |v_d|.$$

De la même manière, on a pour tout $\theta \in U_\rho$, $v_1, \ldots, v_d$ dans $\mathbb{C}^d$ et $i \geq 0$,

$$\left| \frac{1}{i!} \nabla_a^i H_1(b^*, \theta)(v_1, \ldots, v_d) \right| \leq \frac{M}{r^i} |v_1| \cdots |v_d|.$$

On en déduit donc que

$$\begin{aligned} \|K_j\|_{(j-1)} \leq \sum_{i=2}^{j} \sum_{k_1 + \ldots + k_i = j} \frac{M}{r^i} \|\nabla_\theta S_{k_1}\|_{(j-1)}, \ldots, \|\nabla_\theta S_{k_i}\|_{(j-1)} \\ + \sum_{i=1}^{j-1} \sum_{k_1 + \ldots + k_i = j-1} \frac{M}{r^i} \|\nabla_\theta S_{k_1}\|_{(j-1)}, \ldots, \|\nabla_\theta S_{k_i}\|_{(j-1)}, \end{aligned}$$

[3]L'ensemble $B_r(b^*)$ désigne l'ensemble des $z \in \mathbb{C}^d$ tels que $|z - b^*| < r$ où $|\cdot|$ désigne la norme $\sum_{i=1}^d |z_i|$ sur $\mathbb{C}^d$

et en utilisant la borne (6.2.9) et le fait que pour tout $k \leq j-1$, on a $\|\cdot\|_{(j-1)} \leq \|\cdot\|_{(k)}$, puis en divisant par r, on en déduit donc que

$$\frac{1}{r}\|\nabla_\theta S_j\|_{(j)} \leq \frac{M\kappa_1}{r\delta^\alpha}\sum_{i=2}^{j}\sum_{k_1+\ldots+k_i=j}\frac{1}{r^i}\|\nabla_\theta S_{k_1}\|_{(k_1)},\ldots,\|\nabla_\theta S_{k_i}\|_{(k_i)}$$
$$+\frac{M\kappa_1}{r\delta^\alpha}\sum_{i=1}^{j-1}\sum_{k_1+\ldots+k_i=j-1}\frac{1}{r^i}\|\nabla_\theta S_{k_1}\|_{(k_1)},\ldots,\|\nabla_\theta S_{k_i}\|_{(k_i)}.$$

Posons $\mu = M\kappa_1\delta^{-\alpha}r^{-1}$, et considérons la suite de nombres positifs β_i définis par $\beta_1 = \mu$ et pour tout $j \geq 2$,

$$\beta_j = \mu\sum_{i=2}^{j}\sum_{k_1+\ldots+k_i=j}\beta_{k_1}\cdots\beta_{k_i} + \mu\sum_{i=1}^{j-1}\sum_{k_1+\ldots+k_i=j-1}\beta_{k_1}\cdots\beta_{k_i}. \tag{6.2.10}$$

Alors puisqu'on a

$$\frac{1}{r}\|\nabla_\theta S_1\|_{(1)} \leq \kappa_1\delta^{-\alpha}r^{-1}\|K_1\|_{(0)} = \kappa_1\delta^{-\alpha}r^{-1}\|H_1\|_{(0)} \leq \frac{\kappa_1}{\delta^\alpha}\frac{M}{r} = \mu,$$

on en déduit par récurrence que pour tout $j \geq 1$, on a

$$\frac{1}{r}\|\nabla_\theta S_j\|_{(j)} \leq \beta_j.$$

Considérons alors la fonction génératrice $b(\zeta) = \sum_{j=1}^{\infty}\beta_j\zeta^j$. L'équation (6.2.10) montre que

$$\sum_{j=2}^{\infty}\zeta^j\beta_j = \mu\sum_{j=2}^{\infty}\zeta^j\sum_{i=2}^{j}\sum_{k_1+\ldots+k_i=j}\beta_{k_1}\cdots\beta_{k_i} + \mu\zeta\sum_{j=1}^{\infty}\zeta^j\sum_{i=1}^{j}\sum_{k_1+\ldots+k_i=j}\beta_{k_1}\cdots\beta_{k_i},$$

ce qu'on écrit encore

$$\sum_{j=2}^{\infty}\zeta^j\beta_j = -\mu\sum_{j=1}^{\infty}\zeta^j\beta_j + \mu\sum_{j=1}^{\infty}\zeta^j\sum_{i=1}^{j}\sum_{k_1+\ldots+k_i=j}\beta_{k_1}\cdots\beta_{k_i} + \mu\zeta\sum_{j=1}^{\infty}\zeta^j\sum_{i=1}^{j}\sum_{k_1+\ldots+k_i=j}\beta_{k_1}\cdots\beta_{k_i}. \tag{6.2.11}$$

Maintenant, on calcule que

$$\frac{1}{1-b(\zeta)} = \sum_{i=0}^{\infty}b(\zeta)^i = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\zeta^j\sum_{k_1+\ldots+k_i=j}\beta_{k_1}\cdots\beta_{k_i} = 1 + \sum_{j=1}^{\infty}\zeta^j\sum_{i=1}^{j}\sum_{k_1+\ldots+k_i=j}\beta_{k_1}\cdots\beta_{k_i}.$$

Ainsi, l'équation (6.2.11) s'écrit-elle encore

$$b(\zeta) - \mu\zeta = -\mu b(\zeta) + \mu(\zeta+1)\left(\frac{1}{1-b(\zeta)} - 1\right).$$

Cette équation est une équation du second degré en $b = b(\zeta)$ et s'écrit

$$(\mu+1)b^2 - b + \mu\zeta = 0.$$

Compte tenu de la condition $\beta_1 = \mu$, cette équation se résout formellement en écrivant :

$$b(\zeta) = \frac{1}{2}\frac{1}{\mu+1} - \sqrt{\frac{1}{4(\mu+1)^2} - \frac{\mu}{\mu+1}\zeta}.$$

Or cette fonction est analytique dans le disque $|\zeta| < 1/(4\mu(\mu+1))$, et sur ce disque, on a

$$|b(\zeta)| \leq \frac{3}{2}\frac{1}{\mu+1}.$$

Les estimations de Cauchy appliquées à la fonction analytique $b(\zeta)$ montrent alors que

$$\forall\, j \geq 1, \quad \beta_j \leq \frac{3}{2}\frac{1}{\mu+1}(4\mu(\mu+1))^j.$$

On trouve donc

$$\forall\, j \geq 1, \quad \|\nabla_\theta S_j\|_{(j)} \leq r\beta_j \leq \frac{3}{2} r 4^j \mu^j (\mu+1)^j.$$

En remplaçant alors μ par sa valeur, on trouve

$$\forall\, j \geq 1, \quad \|\nabla_\theta S_j\|_{(j)} \leq \frac{3}{2} r 4^j (M\kappa_1\delta^{-\alpha}r^{-1})^j (M\kappa_1\delta^{-\alpha}r^{-1}+1)^j.$$

Si $M\kappa_1\delta^{-\alpha}r^{-1} \geq 1$, on trouve donc

$$\forall\, j \geq 1, \quad \|\nabla_\theta S_j\|_{(j)} \leq \frac{3}{2} r 8^j (M\kappa_1\delta^{-\alpha}r^{-1})^{2j}.$$

En écrivant cette inégalité pour $j = J$, et en remplaçant alors δ par sa valeur $\rho/2J$, on trouve

$$\|\nabla_\theta S_J\|_{(J)} \leq \frac{3}{2} r 8^j (M\kappa_1\rho^{-\alpha}r^{-1}2^\alpha J^\alpha)^{2j},$$

ce qui montre le résultat car

$$\|\nabla_\theta S_J\|_{(J)} = \|\nabla_\theta S_J\|_{\rho-J\delta} = \|\nabla_\theta S_J\|_{\rho/2}.$$

Notons que dans le cas inintéressant où $M\kappa_1\delta^{-\alpha}r^{-1} \leq 1$, qui n'a jamais lieu pour des J assez grand, on trouve immédiatement

$$\forall\, j \geq 1, \quad \|\nabla_\theta S_j\|_{(j)} \leq \frac{3}{2} r 8^j.$$

□

Ainsi, si b^* satisfait la condition diophantienne, et si les données sont analytiques en ce point, on peut résoudre (6.2.4) et trouver des estimations pour la suite de fonctions S_j.

6.3 Estimations d'erreurs

Considérons maintenant un entier N, et considérons la fonction génératrice donnée par

$$S(b^*,\theta) = b^* \cdot \theta + \varepsilon S_1(b^*,\theta) + \ldots + \varepsilon^{N-1} S_{N-1}(b^*,\theta),$$

qui est bien définie dès que b^* satisfait la condition diophantienne (6.2.5). On peut donc considérer le polynôme en ε

$$a(\varepsilon) = b^* + \sum_{k=1}^{N-1} \varepsilon^k \nabla_\theta S_k(b^*,\theta),$$

et le hamiltonien

$$\mathcal{H}(\varepsilon)(b^*,\theta) = H_\varepsilon(a(\varepsilon),\theta),$$

où $H_\varepsilon(a,\theta) = H_0(a) + \varepsilon H_1(a,\theta)$.

Il est clair en développant en puissances de ε, que ce hamiltonien s'écrit

$$\mathcal{H}(\varepsilon)(b^*,\theta) = H_0(b^*) + \sum_{j=1}^{N-1} \varepsilon^j \langle K_j\rangle(b^*) + \varepsilon^N R_N(b^*,\theta),$$

où

$$R_N(b^*,\theta) = \sum_{i=2}^{N} \sum_{k_1+\ldots+k_i=N} \frac{1}{i!}\nabla_a^i H_0\left(Q_{k_1},\ldots,Q_{k_i}\right) + \sum_{i=1}^{N-1} \sum_{k_1+\ldots+k_i=N-1} \frac{1}{i!}\nabla_a^i H_1\left(Q_{k_1},\ldots,Q_{k_i}\right),$$

où Q_k est défini par

$$Q_k = \nabla_\theta S_k + \varepsilon\nabla_\theta S_{k+1} + \ldots + \varepsilon^{N-k-1}\nabla_\theta S_{N-1}.$$

6.3.1 Estimation du reste

En utilisant les estimations précédentes, on montre alors le résultat suivant :

Proposition 3 *Sous les hypothèses de la Proposition 2 avec $r \leq 1$, alors si N satisfait la condition*

$$1 \leq C_1 N^{2\alpha} \leq \frac{1}{2\varepsilon}, \tag{6.3.1}$$

on a l'estimation

$$\|R_N(b^*,\cdot)\|_{\rho/2} \leq 3Mr\left(\frac{4C_1}{r}N^{2\alpha}\right)^N,$$

où M est défini en (6.2.8).

Preuve : En utilisant la Proposition 2 et l'expression de Q_k, on trouve que

$$\|Q_k(b^*,\cdot)\|_{\rho/2} \leq \sum_{j=k}^{N-1} \varepsilon^{j-k} C_0 (C_1 j^{2\alpha})^j.$$

Sous l'hypothèse (6.3.1), qui s'écrit encore $\varepsilon \leq 1/(2C_1N^{2\alpha})$, on trouve

$$\|Q_k(b^*,\cdot)\|_{\rho/2} \leq C_0 \sum_{j=k}^{N-1} 2^{k-j} C_1^{k-j} (N^{2\alpha})^{k-j} C_1^j j^{2\alpha j} = C_0 (C_1 N^{2\alpha})^k \sum_{j=k}^{N-1} 2^{k-j}\left(\frac{j}{N}\right)^{2\alpha j},$$

d'où

$$\|Q_k(b^*,\cdot)\|_{\rho/2} \leq C_0(C_1N^{2\alpha})^k \sum_{j=k}^{N-1} 2^{k-j} \leq 2C_0(C_1N^{2\alpha})^k.$$

On trouve donc en utilisant les estimations de Cauchy que

$$\|R_N(b^*,\theta)\|_{\rho/2} \leq \sum_{i=2}^{N} \sum_{k_1+\ldots+k_i=N} \frac{M}{r^i} 2C_0(C_1N^{2\alpha})^N + \sum_{i=1}^{N-1} \sum_{k_1+\ldots+k_i=N-1} \frac{M}{r^i} 2C_0 (C_1N^{2\alpha})^{N-1},$$

d'où

$$\begin{aligned}
\|R_N(b^*,\theta)\|_{\rho/2} &\leq 2MC_0\left(\frac{C_1}{r}N^{2\alpha}\right)^N\left(\sum_{i=2}^{N}\binom{N+i-1}{i}+\sum_{i=1}^{N-1}\binom{N+i-2}{i}\right)\\
&\leq 2MC_0\left(\frac{C_1}{r}N^{2\alpha}\right)^N\left(2\sum_{i=0}^{N}\binom{2N-1}{i}\right)\\
&\leq 2MC_0\left(\frac{4C_1}{r}N^{2\alpha}\right)^N,
\end{aligned}$$

d'où le résultat. □

6.3.2 Résultats pour une perturbation de type polynôme trigonométrique

La situation se résume ainsi : si on se fixe un b^* satisfaisant la condition diophantienne, on sait construire pour tout entier N une fonction génératrice

$$S(b^*,\theta)=b^*\cdot\theta+\varepsilon S_1(b^*,\theta)+\ldots+\varepsilon^{N-1}S_{N-1}(b^*,\theta),$$

telle que si on définit le polynôme en ε

$$a(\varepsilon)=b^*+\sum_{k=1}^{N-1}\varepsilon^k\nabla_\theta S_k(b^*,\theta),$$

alors le hamiltonien $\mathcal{H}(\varepsilon)(b^*,\theta)=H_\varepsilon(a(\varepsilon),\theta)$ s'écrit

$$\mathcal{H}(\varepsilon)(b^*,\theta)=H_0(b^*)+\sum_{j=1}^{N-1}\varepsilon^j\langle K_j\rangle(b^*)+\varepsilon^N R_N(b^*,\theta), \tag{6.3.2}$$

où le reste s'estime comme précédemment.

Toutefois, la fonction S n'est pas une fonction génératrice au sens où elle n'est définie qu'en b^*. L'idée principale se résume ainsi : on suppose que la perturbation H_1 est un polynôme trigonométrique de dégré m en θ. Par un argument de perturbation, on montre alors que la condition diophantienne est satisfaite pour des b proches de b^* et des k bornés, les bornes dépendant de N et m. On montre alors qu'on peut obtenir obtenir l'existence d'une "vraie" fonction génératrice $S(b,\theta)$ et un hamiltonien $\mathcal{H}(\varepsilon)(b,\theta)$ de la forme (6.3.2), tous analytiques en b dans un voisinage de b^*. Dans la sous-section suivante, on conclut en approchant H_1 par un polynôme trigonométrique bien choisi.

Le lemme suivant montre que la condition diophantienne est satisfaite dans un voisinage de b^* à condition de se restreindre à un nombre fini de fréquence. Cette idée est déjà présente dans les travaux de Poincaré. Elle a été pleinement réalisée dans la démonstration du théorème KAM d'Arnold en 1963.

Lemme 9 (Ultraviolet cut-off) *Supposons que H_0 soit analytique dans un voisinage de b^*. Soit $\omega(b)=\nabla_\theta H_0(b)$. Supposons que b^* satisfait la condition diophantienne* (6.2.5), *et soit*

$$c^*=\frac{\gamma}{2(\max\|\omega'\|)}$$

où $\|\omega'\|$ désigne la norme de la dérivée de ω dans le domaine $B_R(b^*)$. Alors on a pour tout entier m,

$$|k\cdot\omega(b)|\geq\frac{\gamma}{2}|k|^{-\nu}\quad\textit{pour}\quad\|b-b^*\|\leq\frac{c^*}{m^{\nu+1}}\quad\textit{et}\quad|k|\leq m.$$

Preuve : Pour tout $k \in \mathbb{Z}^d$, on a

$$|k \cdot (\omega(b) - \omega(b^*))| \leq |k|(\max \|\omega'\|)\|b - b^*\| ,$$

d'où si $\|b - b^*\| \leq c^* m^{-\nu-1}$,

$$\begin{aligned} |k \cdot \omega(b)| &\geq |k \cdot \omega(b^*)| - |k|(\max \|\omega'\|)\|b - b^*\| \\ &\geq \gamma |k|^{-\nu} - (\max \|\omega'\|) c^* m^{-\nu}, \end{aligned}$$

où $\max \|\omega'\|$ désigne le maximum de la dérivée de ω dans un voisinage assez grand de b^*, indépendant de m. On en déduit que pour $c^* = \gamma/(2(\max \|\omega'\|))$, on a

$$|k \cdot \omega(b)| \geq \frac{\gamma}{2}|k|^{-\nu} \quad \text{pour} \quad \|b - b^*\| \leq c^* m^{-\nu-1} \quad \text{et} \quad |k| \leq m,$$

Ce qui montre le résultat. □

Grâce à ce résultat, on peut étendre les résultats et estimations précédents à des voisinages de b^* :

Proposition 4 *Supposons que H_1 s'écrit sous la forme*

$$H_1(a, \theta) = \sum_{k \in \mathbb{Z}^d,\, |k| \leq m} h_k(a) e^{ik \cdot \theta}$$

où m est un entier, et où les $h_k(a)$ sont des fonctions analytiques sur D. Supposons de plus que b^* satisfait la condition diophantienne (6.2.5). *Alors pour tout $j \geq 1$ il existe des polynômes trigonométriques $S_j(b, \theta)$ de degrés jm, solutions des équations* (6.2.4) *qui sont analytiques pour (b, θ) tels que $\|b - b^*\| \leq c^*(jm)^{-\nu-1}$ et $\theta \in U_\rho$, où c^* est une constante ne dépendant que de γ et de bornes sur H_0. De plus, on a les estimations*

$$\|\nabla_\theta S_j(b, \theta)\|_{\rho/2} \leq C_0 (4 C_1 j^{2\alpha})^j, \quad \textit{pour} \quad \|b - b^*\| \leq c^*(jm)^{-\nu-1}. \tag{6.3.3}$$

Pour un N fixé, le reste $R_N(b, \theta)$ est aussi analytique pour (b, θ) tels que $\|b - b^*\| \leq c^*(Nm)^{-\nu-1}$ et $\theta \in U_\rho$, et on a l'estimation

$$\|R_N(b, \theta)\|_{\rho/2} \leq 2 C_0 (C_2 N^{2\alpha})^N, \quad \textit{pour} \quad \|b - b^*\| \leq c^*(Nm)^{-\nu-1}, \tag{6.3.4}$$

où $C_2 = 16 C_1 / r$, pour des N tels que

$$1 \leq C_2 N^{2\alpha} \leq \frac{1}{2\varepsilon}.$$

Preuve : La preuve de ce résultat se fait en reprenant en grande partie les calculs précédents. Dans la proposition 1, supposons que G soit un polynôme trigonométrique de dégré jm, c'est-à-dire qu'on a avec les notations de la preuve $g_k(b) = 0$ pour $|k| > jm$. Supposons de plus que les fonctions $g_k(b)$ son définies et analytiques pour b satisfaisant

$$\|b - b^*\| \leq c^*(jm)^{-\nu-1}.$$

Alors pour b fixé, on voudrait définir les coefficients $f_k(b)$ par l'équation

$$\forall\, k, \quad |k| \leq jm, \quad f_k(b) = -\frac{g_k(b)}{ik \cdot \omega(b)}.$$

Or d'après le lemme précédente, on a

$$|k \cdot \omega(b)| \geq \frac{\gamma}{2}|k|^{-\nu} \quad \text{pour} \quad \|b - b^*\| \leq c^*(jm)^{-\nu-1} \quad \text{et} \quad |k| \leq jm.$$

Ainsi, on peut définir la fonction

$$F(b,\theta) = \sum_{k \in \mathbb{Z}^d, |k| \leq jm} -\frac{g_k(b)}{ik \cdot \omega(b)} e^{ik\cdot\theta}$$

pour tout b tel que $\|b - b^*\| \leq c^*(jm)^{-\nu-1}$ et $\theta \in U_\rho$. Il est alors facile de vérifier que cette fonction est analytique sur cet ensemble, et satisfait des estimations similaires à celle de la Proposition 1 avec γ changé en $\gamma/2$ (cette transformation est à l'origine du changement de C_1 en $4C_1$ dans les estimations énoncées ci-dessus).

Considérons maintenant le processus (6.2.4). La démonstration de la Proposition 2 montre par récurrence que les K_{j+1} sont des polynômes trigonométriques de degré au plus $(j+1)m$ analytiques pour (b,θ) tels que $\theta \in U_\rho$ et $\|b - b^*\| \leq c^*(jm)^{-\nu-1}$ donc a fortiori pour $\|b-b^*\| \leq c^*((j+1)m)^{-\nu-1}$.

En suivant la démonstration de la proposition 2, on montre l'estimation (6.3.3), quitte à modifier r et M afin d'utiliser des estimations de type Cauchy sur des voisinages de b^* indépendants[4] de j. L'estimation (6.3.4) se montre de la même manière à partir de la Proposition 3. □

Considérons maintenant la fonction génératrice

$$S(b,\theta) = b \cdot \theta + \varepsilon S_1(b,\theta) + \ldots + \varepsilon^{N-1} S_{N-1}(b,\theta),$$

définie pour $\|b - b^*\| \leq c^*(Nm)^{-\nu-1}$ et $\theta \in U_{\rho/2}$. Cette fonction génératrice définit une transformation symplectique $\Psi : (a,\theta) \mapsto (b,\varphi)$ qui est ε proche de l'identité, au sens où $\Psi = \mathrm{Id} + \mathcal{O}(\varepsilon)$ et Ψ, Ψ^{-1} et leurs dérivées sont toujours bornées indépendamment de ε.

Dans ces nouvelles variables, le hamiltonien s'écrit

$$\mathcal{H}(\varepsilon)(b,\varphi) = H_0(b) + \sum_{j=1}^{N-1} \varepsilon^j \langle K_j \rangle(b) + \varepsilon^N R_N(b, \theta(b,\varphi)),$$

où $\theta(b,\varphi)$ désigne la composante θ du difféomorphisme Ψ^{-1}.

Le système hamiltonien d'origine s'écrit donc, pour $\|b - b^*\| \leq c^*(Nm)^{-\nu-1}$,

$$\dot{b} = -\nabla_\varphi \mathcal{H}(\varepsilon)(b,\varphi) = -\varepsilon^N (\nabla_\varphi \theta(b,\varphi)) \cdot (\nabla_\theta R_N(b, \theta(b,\varphi))$$

et

$$\begin{array}{rcl}\dot{\varphi} & = & \nabla_b \mathcal{H}(\varepsilon)(b,\varphi) \\ & = & \nabla_b H_0(b) + \sum_{j=1}^{N-1} \varepsilon^j \nabla_b \langle K_j \rangle(b) + \varepsilon^N \nabla_b R_N(b, \theta(b,\varphi)) \\ & & +\varepsilon^N (\nabla_b \theta(b,\varphi)) \cdot (\nabla_\theta R_N(b, \theta(b,\varphi)).\end{array}$$

En utilisant le fait que $\theta(b,\varphi) = \varphi + \mathcal{O}(\varepsilon)$, et en utilisant les estimations précédentes associées aux estimations de Cauchy, on trouve que

$$\begin{array}{rcl}\dot{b} & = & \mathcal{O}(\varepsilon^N (C_2 N^{2\alpha})^N), \\ \dot{\varphi} & = & \omega_{\varepsilon,N}(b) + \mathcal{O}((Nm)^{\nu+1} \cdot \varepsilon^N (C_2 N^{2\alpha})^N),\end{array} \tag{6.3.5}$$

[4] Au sens strict, les constantes C_0 et C_1 sont les mêmes que précédemment, mais où M est un maximum pris sur une boule de rayon R tandis que r est modifié en $R - r > 0$. Ceci est nécessaire pour obtenir des estimations de Cauchy dans un voisinage de b^*.

où on a posé

$$\omega_{\varepsilon,N}(b) = \nabla_b H_0(b) + \sum_{j=1}^{N-1} \varepsilon^j \nabla_b \langle K_j \rangle (b).$$

Notons que dans les équations (6.3.5), $\nabla_\varphi \theta$ désigne la dérivée de la composante θ de Ψ^{-1} par rapport à φ, et donc est borné indépendamment de ε.

6.3.3 Un théorème de perturbation

On fixe maintenant un entier N, et on suppose que la perturbation $H_1(a,\theta)$ est analytique. On suppose de plus toujours que ε est assez petit pour pouvoir appliquer la proposition 3. Si on cherche à approcher H_1 à ε^N près, il faut prendre m de l'ordre de $|\log \varepsilon^N|$. Plus précisemment, on à le résultat suivant :

Lemme 10 *Supposons $H_1(b,\theta)$ soit analytique pour des (b,θ) tels que b soit proche de b^* et $\theta \in U_\rho$. Soit*

$$H_1(b,\theta) = \sum_{k\in\mathbb{Z}^d} h_k(b) e^{ik\cdot\omega(b)}$$

le développement en série de Fourier de $H_1(b,\theta)$ dans ce domaine. Pour un entier m, on définit la série de Fourier tronquée

$$H_1^{(m)}(b,\theta) = \sum_{k\in\mathbb{Z}^d, |k|\le m} h_k(b) e^{ik\cdot\omega(b)}.$$

Alors si on a $m = E((-2N\log\varepsilon)/\rho)$, alors[5] *il existe un $C > 0$ tel que pour tout b on ait*

$$\|H_1(b,\theta) - H_1^{(m)}(b,\theta)\|_{\rho/2} \le C\varepsilon^N.$$

Preuve : La différence s'écrit

$$H_1(b,\theta) - H_1^{(m)}(b,\theta) = \sum_{k\in\mathbb{Z}^d, |k|>m} h_k(b) e^{ik\cdot\omega(b)}.$$

Or on sait qu'il existe un constante Γ telle que pour tout k et pour tout b dans un voisinage assez petit de b^*, on ait

$$|h_k(b)| \le \Gamma e^{-\rho|k|}.$$

On en déduit que pour $\theta \in U_{\rho/2}$, on a

$$|H_1(b,\theta) - H_1^{(m)}(b,\theta)| \le \Gamma \sum_{k\in\mathbb{Z}^d, |k|>m} e^{-\rho/2|k|}.$$

Par des calculs similaires à ceux de la proposition 1, on trouve donc que

$$|H_1(b,\theta) - H_1^{(m)}(b,\theta)| \le Ce^{-(m+1)\rho/2}.$$

Si maintenant $m = E((-2N\log\varepsilon)/\rho)$, ($\varepsilon$ est toujours supposé plus petit que 1), on a $m+1 \ge (-2N\log\varepsilon)/\rho$, et on trouve donc

$$|H_1(b,\theta) - H_1^{(m)}(b,\theta)| \le Ce^{N\log\varepsilon} = C\varepsilon^N,$$

[5] $E(x)$ désigne la partie entière de x.

d'où le résultat. □

Soit maintenant N fixé et $m = E((-2N\log\varepsilon)/\rho)$ sous les hypothèses du lemme précédent. On peut donc écrire que $H_1 = H_1^{(m)} + \mathcal{O}(\varepsilon^N)$, et utiliser les équations (6.3.5). L'entier N étant considéré fixe, ces équations s'écrivent alors, $\|b - b^*\| \leq c|\log\varepsilon|^{-\nu-1}$,

$$\begin{aligned} \dot{b} &= \mathcal{O}(\varepsilon^N), \\ \dot{\varphi} &= \omega_{\varepsilon,N}(b) + \mathcal{O}(|\log\varepsilon|)^{\nu+1}\varepsilon^N). \end{aligned} \tag{6.3.6}$$

On en déduit le résultat suivant :

Théorème 17 *Supposons que $H_0(b)$ soit analytique au voisinage d'un point b^* satisfaisant la condition diophantienne* (6.2.5)*. Supposons de plus que $H_1(b,\theta)$ soit analytique pour des b dans un voisinage de b^* et $\theta \in U_\rho$ pour $\rho > 0$. Alors pour tout N fixé, il existe $\varepsilon_0 > 0$, et des constantes c et C, tels qu'il existe un changement de variable symplectique $\Psi : (a,\theta) \mapsto (b,\varphi)$ défini (ainsi que son inverse) sur l'ensemble*

$$\{(a,\theta)\,;\, \|a - b^*\| \leq c|\log\varepsilon|^{-\nu-1}, \varphi \in U_{\rho/4}\} \tag{6.3.7}$$

proche de l'identité au sens où

$$\|(a,\theta) - \Psi(a,\theta)\| \leq C\varepsilon$$

pour (a,θ) dans l'ensemble (6.3.7) *(et une relation similaire pour (b,φ) et Ψ^{-1}), tel que le système hamiltonien associé au hamiltonien* (6.2.1) *satisfait la propriété suivante : toute solution de ce système $(b(t),\varphi(t))$ dans les variables (b,φ) telle que*

$$\|b(0) - b^*\| \leq c_2|\log\varepsilon|^{-\nu-1}$$

satisfait

$$\|b(t) - b(0)\| \leq Ct\varepsilon^N \quad \textit{pour} \quad t \leq \varepsilon^{-N+1} \tag{6.3.8}$$

et

$$\|\varphi(t) - t\omega_{\varepsilon,N}(b(0)) - \varphi(0)\| \leq C(t^2 + t|\log\varepsilon|^{\nu+1})\varepsilon^N \quad \textit{pour} \quad t^2 \leq \varepsilon^{-N+1}. \tag{6.3.9}$$

Preuve : L'existence du difféomorphisme symplectique Ψ est une conséquence de ce qui précède, et de ce que Ψ est une perturbation de l'identité. Les équations (6.3.6) sont valables tant que (b,φ) reste dans le domaine (6.3.7).

L'estimation (6.3.8) est alors claire, si ε_0 est assez petit pour que $C\varepsilon_0 \leq c_2|\log\varepsilon|^{-\nu-1}$. L'estimation (6.3.9) est alors une conséquence de (6.3.6) et (6.3.7) en utilisant un développement de Taylor de $\omega_{\varepsilon,N}(b)$ autour de $b(0)$ (les dérivées de $\omega_{\varepsilon,N}(b)$ sont bornées indépendamment de ε). De plus, φ va bien rester dans $U_{\rho/4}$ pour ε_0 assez petit. Ceci montre le théorème. □

On peut encore raffiner ce théorème en optimisant sur le N pour obtenir des estimations avec des termes d'erreurs exponentiels, sur des temps exponentiellement long. En effet, en prenant N tel que $C_2N^{2\alpha} = 1/(e\varepsilon^\beta)$, alors si $\beta \leq 1$, on est dans le cadre de la proposition 3 (voir l'équation (6.3.1)) et les équations (6.3.5) s'écrivent d'une part

$$\dot{b} = \mathcal{O}\left(\exp((C_2e)^{-2\alpha}\varepsilon^{-\beta/2\alpha}((1-\beta)\log\varepsilon - 1)))\right)$$

en utilisant le fait que $N = (C_2e)^{-1/2\alpha}\varepsilon^{-\beta/2\alpha}$,

$$\dot{\varphi} = \omega_{\varepsilon,N}(b) + \mathcal{O}\left(\varepsilon^{-2\beta}\exp((C_2e)^{-1/2\alpha}\varepsilon^{-\beta/2\alpha}((1-\beta)\log\varepsilon - 1)))\right).$$

On trouve finalement que pour $\varepsilon < \varepsilon_0$ assez petit, on a

$$\begin{aligned} \dot{b} &= \mathcal{O}\left(\exp(-c\varepsilon^{-\beta/2\alpha})\right) \\ \dot{\varphi} &= \omega_{\varepsilon,N}(b) + \mathcal{O}\left(\varepsilon^{-2\beta}\exp(-c\varepsilon^{-\beta/2\alpha})\right) \end{aligned}$$

pour $\|b - b^*\| \leq C\varepsilon^{2\beta}$, où la constante c vaut $(C_2 e)^{-1/2\alpha}$. On en déduit des estimations similaires à (6.3.8) et (6.3.9).

6.4 Applications aux schémas symplectiques

On considère maintenant un système hamiltonien

$$\dot{p} = -\nabla_q H(p,q) \quad \text{et} \quad \dot{q} = \nabla_p H(p,q) \tag{6.4.1}$$

auquel on applique un schéma symplectique de pas h fixé et d'ordre $r \geq 2$. On note (p_n, q_n) la solution numérique. On suppose que ce système est complètement intégrable au sens du théorème d'Arnold-Liouville. On suppose de plus que H est réel analytique. On note $(p(t), q(t))$ la solution exacte de (6.4.1).

Dans la suite, on note

$$(a, \theta) = (I(p,q), \Theta(p,q))$$

des variables actions-angles associées à ce problème. Les fonctions $I(p,q)$ sont alors des intégrales premières du système. Dans les variables actions-angles, le hamiltonien s'écrit $\mathcal{H}(a) = H(p,q)$. On note alors

$$\omega(a) = \nabla_a \mathcal{H}(a)$$

la fréquence du système. Dans les variables (a, θ), les solutions du système (6.4.1) s'écrivent

$$a(t) = a_0 \quad \text{et} \quad \theta(t) = \omega(a(0))t + \theta_0.$$

Le résultat suivant est une conséquence du paragraphe précédent :

Théorème 18 *Sous les hypothèses précédentes, supposons que $\omega(a^*)$ satisfait la condition diophantienne* (6.2.5)*. Alors il existe des constantes C, c et h_0 dépendant de la dimension d, des constantes γ, ν apparaissant dans la condition diophantienne, et de bornes du hamiltonien $H(p,q)$ dans un voisinage complexe du tore $\{(p,q)\,;\, I(p,q) = a^*\}$ telles que pour tout $h \leq h_0$, toute solution numérique (p_n, q_n) telle que*

$$\|I(p_0,q_0) - a^*\| \leq c|\log h|^{-\nu-1}$$

vérifie pour tout n tel que $n \leq h^{-r-1}$,

$$\left.\begin{aligned} \|(p_n,q_n) - (p(t),q(t))\| &\leq Cth^r \\ \|I(p_n,q_n) - I(p_0,q_0)\| &\leq Ch^r \end{aligned}\right. \qquad \text{pour} \quad t = nh \leq h^{-r}.$$

Preuve : D'après le Théorème 11, les termes f_j de l'équation modifiée sont de la forme $J^{-1}\nabla H_j$. On peut donc définir pour tout l le Hamiltonien modifié

$$\tilde{H} = H + h^r H_{r+1} + \ldots + h^l H_{l+1}.$$

Dans la suite, on note $(\tilde{p}(t), \tilde{q}(t))$ la solution exacte associée à ce Hamiltonien avec valeur initiale (p_0, q_0). On commence par estimer la différence entre $(\tilde{p}(t), \tilde{q}(t))$ et la solution exacte $(p(t), q(t))$.

On estime ensuite la différence entre la solution de l'équation modifiée tronquée et de la solution numérique.
(a) Dans les variables actions-angles, le hamiltonien modifié $\tilde{H}(p,q) = \tilde{\mathcal{H}}(a,\theta)$ s'écrit

$$\tilde{\mathcal{H}}(a,\theta) = \mathcal{H}(a) + h^r \mathcal{H}_h(a,\theta),$$

où le hamiltonien $\mathcal{H}_h(a,\theta)$ est uniformément borné en h (et en l) dans un voisinage complexe du tore $a^* \times \mathbb{T}^d$ indépendant de h. On peut alors appliquer le théorème 15 avec $\varepsilon = h^r$. On trouve qu'il existe un changement de variable symplectique $(a,\theta) \mapsto (b,\varphi)$, h^r proche de l'identité, tel que dans ces nouvelles coordonnées, le système associé à l'équation modifiée s'écrive

$$\begin{array}{rcl} \tilde{b}(t) & = & \tilde{b}(0) + \mathcal{O}(th^{rN}), \\ \tilde{\varphi}(t) & = & \tilde{\varphi}(0) + t\omega_h(\tilde{b}(0)) + \mathcal{O}(th^{rN}|\log h|^{\nu+1} + t^2 h^{rN}) \end{array} \quad \text{pour} \quad t \leq h^{-rN+r}. \tag{6.4.2}$$

avec $\omega_h(b) = \omega(b) + \mathcal{O}(h^r)$. Les constantes présentes dans les $\mathcal{O}$ ne dépendent pas de h, ni de $t \leq h^{-r}$, ni de $(\tilde{b}(0), \tilde{\varphi}(0))$ tel que $|\tilde{b}(0) - a^*| \leq c|\log h|^{-\nu-1}$.

Comme le changement de variable symplectique Ψ est h^r proche de l'identité, on trouve que dans les variables actions-angles, la solution des équations du champ modifié s'écrit

$$\begin{array}{rcl} \tilde{a}(t) & = & \tilde{a}(0) + \mathcal{O}(h^r + th^{rN}), \\ \tilde{\theta}(t) & = & \tilde{\theta}(0) + t\omega(\tilde{a}(0)) + te_h + \mathcal{O}(h^r + th^{rN}|\log h|^{\nu+1} + t^2 h^{rN}) \end{array} \quad \text{pour} \quad t \leq h^{-rN+r},$$

où $e_h = \omega_h(\tilde{b}(0)) - \omega(\tilde{a}(0)) = \mathcal{O}(h^r)$. En prenant $N \geq 4$, on trouve donc

$$\begin{array}{rcl} \tilde{a}(t) & = & \tilde{a}(0) + \mathcal{O}(h^r), \\ \tilde{\theta}(t) & = & \tilde{\theta}(0) + t\omega(\tilde{a}(0)) + \mathcal{O}(th^r) \end{array} \quad \text{pour} \quad t \leq h^{-r}.$$

En écrivant alors les équations pour le champ exact, puis en recomposant par le difféomorphisme symplectique ψ donné par le théorème d'Arnold-Liouville, on trouve que

$$\begin{array}{l} (\tilde{p}(t), \tilde{q}(t)) - (p(t), q(t)) = \mathcal{O}(th^r) \\ I(\tilde{p}(t), \tilde{q}(t)) - I(p(t), q(t)) = \mathcal{O}(h^r) \end{array} \quad \text{pour} \quad t \leq h^{-r},$$

ce qui montre le résultat pour le champ modifié.
(b) Il reste maintenant à borner la différence entre la solution numérique (p_n, q_n) et la solution exacte de l'équation modifiée $(\tilde{p}(t), \tilde{q}(t))$. Or dans les variables (b,φ), si $l = 2r$, le schéma numérique s'écrit d'après l'équation (6.4.2) et la définition du champ modifié, pour $t = nh \leq h^{-r}$ avec $N \geq 4$,

$$b_{n+1} = b_n + \mathcal{O}(h^{l+1} + th^{rN}) = b_n + \mathcal{O}(h^{2r+1})$$

et de même

$$\varphi_{n+1} = \varphi_n + h\omega_h(b_n) + \mathcal{O}(h^{2r+1}).$$

On en déduit que pour $t = nh \leq h^{-r}$,

$$b_n = \tilde{b}(t) + \mathcal{O}(th^{2r}) \quad \text{et} \quad \varphi_n = \tilde{\varphi}(t) + \mathcal{O}(t^2 h^{2r}).$$

En prenant $t \leq h^{-r}$, on a que $th^{2r} \leq h^r$. On trouve donc en composant par deux changement de variables symplectiques (ψ et Ψ) que

$$\begin{array}{l} (\tilde{p}(t), \tilde{q}(t)) - (p_n, q_n) = \mathcal{O}(th^r) \\ I(\tilde{p}(t), \tilde{q}(t)) - I(p_n, q_n) = \mathcal{O}(h^r) \end{array} \quad \text{pour} \quad t \leq h^{-r},$$

ce qui montre le résultat. En groupant les parties (a) et (b) de la preuve, on obtient les estimations du théorème. □

En optimisant sur le N, on trouve des estimations pour des temps exponentiellement longs pour la conservation des invariants.

Chapitre 7

Résultats annexes

7.1 Structure de Poisson canonique

On regroupe ici quelques définitions et résultats liés aux crochets de Poisson.

Définition 5 *Le crochet de Poisson (canonique) de F et G fonctions de $C^1(\mathbb{R}^{2n}, \mathbb{R})$ est défini par :*

$$\{F, G\} = \nabla F^T J^{-1} \nabla G. \tag{7.1.1}$$

Lemme 11 *Le crochet de Poisson est bilinéaire, anti-symétrique, et satisfait l'identité suivante de Jacobi :*

$$\{\{F, G\}, H\} + \{\{G, H\}, F\} + \{\{H, F\}, G\} = 0, \tag{7.1.2}$$

où F, G et H sont trois fonctions de $C^1(\mathbb{R}^{2d}, \mathbb{R})$.

Preuve : La bilinéarité et l'anti-symétrie sont immédiates. L'identité de Jacobi s'écrit :

$$\begin{array}{rl} \{\{F, G\}, H\} + \{\{G, H\}, F\} + \{\{H, F\}, G\} & = \\ (\nabla F)^T J^{-1}(\nabla^2 G) J^{-1}(\nabla H) - (\nabla G)^T J^{-1}(\nabla^2 F) J^{-1}(\nabla H) & \\ +(\nabla G)^T J^{-1}(\nabla^2 H) J^{-1}(\nabla F) - (\nabla H)^T J^{-1}(\nabla^2 G) J^{-1}(\nabla H) & \\ +(\nabla H)^T J^{-1}(\nabla^2 F) J^{-1}(\nabla G) - (\nabla F)^T J^{-1}(\nabla^2 H) J^{-1}(\nabla F) & \\ & = \ 0. \end{array}$$

□

Maintenant, si F est une fonction de $C^1(\mathbb{R}^{2d}, \mathbb{R})$, on peut écrire la dérivée de F le long du flot de $\dot{y} = J^{-1}\nabla H$ comme :

$$\frac{d}{dt} F(\varphi_t(y)) = (\nabla F)(\varphi_t(y))\dot{\varphi}_t(y) = (\nabla F)(\varphi_t(y)) J^{-1}(\nabla H)(\varphi_t(y)) = \{F, H\}(\varphi_t(y)). \tag{7.1.3}$$

Ainsi, F est un invariant du système si et seulement si $\{F, H\} = 0$.

7.2 Crochet de Lie

Définition 6 *Soit f un champ de vecteur ($f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$) et F une fonction de $C^1(\mathbb{R}^n, \mathbb{R}^m)$. L'opérateur* dérivée de Lie *D est défini par :*

$$D = \left(\frac{\partial \cdot}{\partial y}\right) f, \tag{7.2.1}$$

au sens où, si F une fonction de $C^1(\mathbb{R}^n, \mathbb{R})$, on a :

$$\forall\, y \in \mathbb{R}^n,\ DF(y) = \left(\frac{\partial F}{\partial y}(y)\right) f(y). \tag{7.2.2}$$

On peut itérerer l'opérateur D et considérer ses puissances D^k. En fait, on a :

$$\frac{d}{dt^k} F(\varphi_t(y)) = (D^k F)(\varphi_t(y)), \tag{7.2.3}$$

de sorte que le développement en série de Taylor de $F(\varphi_t(y))$ en $t = 0$ s'écrit :

$$F(\varphi_t(y)) = \sum_{k \geq 0} \frac{t^k}{k!}(D^k F)(y) = (\exp(tD)F)(y), \tag{7.2.4}$$

et pour $F(y) = y$:

$$\varphi_t(y) = \sum_{k \geq 0} \frac{t^k}{k!}(D^k Id)(y). \tag{7.2.5}$$

Notons que la série converge si f est analytique.

Considérons maintenant le cas de deux champs de vecteurs $f^{[1]}$ et $f^{[2]}$, fonctions de $C^1(\mathbb{R}^n, \mathbb{R}^n)$. La composition des flots $\varphi_s^{[1]}$ et $\varphi_t^{[2]}$ peut alors s'écrire (formellement) comme :

$$(\varphi_t^{[2]} \circ \varphi_s^{[1]})(y) = (\exp(sD_1)\exp(tD_2)Id)(y). \tag{7.2.6}$$

Les opérateurs apparaissent dans l'ordre inverse car (7.2.6) s'obtient en considérant $F(y) = \varphi_t^{[2]}(y)$ et en développant $F(\varphi_s^{[1]}(y))$. Les séries considérées sont des séries formelles, en général non convergentes, car les opérateurs D_1 et D_2 sont non bornés. Par ailleurs, si D_1 et D_2 ne commutent pas, alors $\exp(sD_1)\exp(tD_2) \neq \exp(sD_1 + tD_2)$. Cependant, on peut formellement procéder à l'identification $\exp(sD_1)\exp(tD_2) = \exp(D(s,t))$ avec

$$D(s,t) = sD_1 + tD_2 + st[D_1, D_2] + \frac{s^2 t}{12}[D_1, [D_1, D_2]] + \frac{st^2}{12}[D_2, [D_2, D_1]] + \dots \tag{7.2.7}$$

où $[D_1, D_2] = D_1 D_2 - D_2 D_1$. Cette identification devient rigoureuse si elle est interprétée comme l'identification de développements de Taylor au voisinage de $s = t = 0$. Les termes suivants du développement sont donnés par la formule Baker-Campbell-Hausdorff (BCH) (voir le livre "Geometric Numerical Integration" pour une présentation complète). La formule (7.2.7) conduit à penser que si $[D_1, D_2] = 0$, alors les flots $\varphi_s^{[1]}$ et $\varphi_t^{[2]}$ commutent. Le résultat est vrai mais ne peut être démontré à partir de (7.2.7) puisque les séries considérées ne sont pas convergentes.

Théorème 19 *Soient $f^{[1]}$ et $f^{[2]}$ deux champs de vecteurs de $C^1(\mathbb{R}^n, \mathbb{R}^n)$. La composition des flots $\varphi_s^{[1]}$ et $\varphi_t^{[2]}$ est commutative si et seulement si*

$$[D_1, D_2] = \left(\frac{\partial\, \cdot}{\partial y}\right)\left(\frac{\partial f^{[2]}}{\partial y} f^{[1]} - \frac{\partial f^{[1]}}{\partial y} f^{[2]}\right) = 0. \tag{7.2.8}$$

Preuve : Quitte à reparamétriser s en multipliant $f^{[1]}$ par une constante (ce qui ne change rien à la nullité du crochet de Lie), on peut considérer $s = t$. On remarque tout d'abord que, d'après la formule (7.2.7), pour tout y de $\mathbb{R}^n$, et pour $h > 0$ suffisament petit :

$$(\varphi_h^{[2]} \circ \varphi_h^{[1]} - \varphi_h^{[2]} \circ \varphi_h^{[1]})(y) = \mathcal{O}(h^3), \tag{7.2.9}$$

où la constante contenue dans $\mathcal{O}$ est obtenue par majoration des dérivées sur un compact contenant $(\varphi_u^{[2]} \circ \varphi_v^{[1]})(y)$ et $(\varphi_v^{[1]} \circ \varphi_u^{[2]})(y)$ pour tout $0 \leq u, v \leq t$. Prenons $h = t/n$. Il vient :

$$\begin{array}{rcl}
\varphi_t^{[2]} \circ \varphi_t^{[1]} & = & \varphi_{t-h}^{[2]} \circ \varphi_h^{[2]} \circ \varphi_h^{[1]} \circ \varphi_{t-h}^{[1]} \\
& = & \varphi_{t-h}^{[2]} \circ \varphi_h^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{t-h}^{[1]} + \mathcal{O}(h^3) \\
& = & \varphi_{t-h}^{[2]} \circ \varphi_{2h}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{t-2h}^{[1]} + 2\mathcal{O}(h^3) \\
& = & \cdots \\
& = & \varphi_{t-h}^{[2]} \circ \varphi_{nh}^{[1]} \circ \varphi_h^{[2]} + n\mathcal{O}(h^3).
\end{array}$$

En répétant l'opération n fois, on obtient :

$$(\varphi_t^{[2]} \circ \varphi_t^{[1]} - \varphi_t^{[1]} \circ \varphi_t^{[2]}) = n^2\mathcal{O}(h^3) = \mathcal{O}(1/n). \tag{7.2.10}$$

Réciproquement, si les flots $\varphi_t^{[1]}$ et $\varphi_s^{[2]}$ commutent pour tout t et tout s, alors :

$$\begin{array}{rcl}
\left.\dfrac{d^2}{dtds}(\varphi_t^{[1]} \circ \varphi_s^{[2]})\right|_{s=t=0} & = & \left.\dfrac{d}{dt}\left(\left.\dfrac{d}{ds}(\varphi_t^{[1]} \circ \varphi_s^{[2]})\right|_{s=0}\right)\right|_{t=0} = \left.\dfrac{d}{dt}(D_2\varphi_t^{[1]})(y)\right|_{t=0} \\
& = & \left.\left(\dfrac{d}{dt}\dfrac{\partial \varphi_t^{[1]}}{\partial y}\right)\right|_{t=0} f^{[2]}(y) = \dfrac{\partial f^{[1]}}{\partial y}(y) f^{[2]}(y).
\end{array}$$

et on conclut en inversant les exposants 1 et 2. □

Remarque 5 *Si $f^{[1]} = J^{-1}\nabla H^{[1]}$ et $f^{[2]} = J^{-1}\nabla H^{[2]}$, alors, pour toute fonction F de $C^1(\mathbb{R}^n, \mathbb{R})$:*

$$\begin{array}{rcl}
D_1F & = & (\nabla F)^T J^{-1}\nabla H^{[1]} = \{F, H^{[1]}\}, \\
D_2F & = & (\nabla F)^T J^{-1}\nabla H^{[2]} = \{F, H^{[2]}\}, \\
D_2D_1F & = & D_2(D_1F) = \{D_1F, H^{[2]}\} = \{\{F, H^{[1]}\}, H^{[2]}\}, \\
D_1D_2F & = & D_1(D_2F) = \{D_2F, H^{[1]}\} = \{\{F, H^{[2]}\}, H^{[1]}\}.
\end{array}$$

En utilisant l'identité de Jacobi, on obtient ainsi :

$$[D_1, D_2](F) = \{F, \{H^{[2]}, H^{[1]}\}\}. \tag{7.2.11}$$

En conclusion, les flots de deux systèmes hamiltoniens commutent ssi $\{H^{[1]}, H^{[2]}\} = 0$.