Cours de culture mathématique

Fondations, Analyse, Géométrie et Applications

Jean Feydy

2017–2018

Version courante adaptée du polycopié de l'année 2016-2017. Au fil des séances de l'automne-hiver 2017, outre un profond remaniement du plan du cours, j'ajouterai remarques introductives et exercices.

Attention : le présent ouvrage n'est pas un manuel autonome, mais un support de référence pour le cours donné à l'ENS au premier semestre. Ce dernier sera donc sensiblement moins formel que les pages qui suivent !

Informations pratiques et remerciements

Contact :

— Adresses mail : `jean.feydy@ens.fr`, `jean.feydy@gmail.com` (utilisez les deux).

— Bureau : sur les toits du DMA, sous la verrière, les lundis et mercredis à coup sûr.

— Page web du cours : `www.math.ens.fr/~feydy/Teaching/index.html`

Inutile d'imprimer ce polycopié, qui est encore loin d'être terminé : j'apporterai à chaque séance une version définitive du chapitre étudié.

L'horaire des séances reste à déterminer. La validation se fera à la présence – huit séances sur douze au minimum – et sur la réponse à un questionnaire en ligne hebdomadaire.

Conseils de lecture On pourra bien entendu lire le présent manuel de bout en bout : c'est ce que nous ferons en classe. Un lecteur pressé pourra toutefois piocher, au gré de ses envies, dans les nombreux sujets abordés. Les chapitres 1, 2 et 4 pourront se lire sans aucun pré-requis ; le chapitre 3 est essentiellement motivé par les résultats de logique énoncés au chapitre 2. Enfin, les parties d'*Analyse* et de *Géométrie* sont indépendantes l'une de l'autre, et seront accessibles à tout lecteur disposant d'un solide bagage de terminale S (nombres complexes, dérivation, intégration).

Remerciements Le présent manuel est un florilège des quatre années que j'ai eu le bonheur de passer à l'École Normale Supérieure (2012-2016). Mes premiers remerciements iront donc tout naturellement à mes professeurs. Blandine Samson (flûte), Martin Hils (logique), Wendelin Werner (analyse complexe), Stéphane Mallat (analyse de Fourier), Ma Xiaoguang (méthode des éléments finis), Haïm Brézis (manuel d'analyse fonctionnelle), Étienne Ghys (manuels de géométrie Riemannienne), Alain Trouvé (espaces de formes) : tous m'ont fait découvrir avec enthousiasme les merveilles d'ingéniosité que je voudrais maintenant partager avec vous.

À Vincent Thouard, Serge Dupont et Olivier Biquard qui, depuis le début, ont su me faire confiance en m'attribuant des tâches d'enseignement toujours plus excitantes : mille mercis !

Enfin, tout ma gratitude va vers Anna Song qui par ses conseils et son regard acéré me pousse chaque jour à aller de l'avant !

Table des matières

1 Introduction **9** ***Séance 1***
1.1 Le jeu de taquin : les étapes du raisonnement mathématique . . . 11 ***Le plan a changé !***
1.1.1 Règles et problème de Loyd . . . 11
1.1.2 Formalisation des règles . . . 11
1.1.3 Signature d'une permutation . . . 16
1.2 Pour aller plus loin : plan du cours . . . 20

I Nombres complexes, géométrie **25**

2 Le corps des nombres complexes **27** ***Séance 2***
2.1 Retour sur les nombres complexes vus au lycée . . . 27
2.1.1 Le présentateur . . . 27
2.1.2 Nombres et transformations . . . 28
2.1.3 La racine carrée de -1 . . . 28
2.1.4 Arithmétique complexe . . . 30
2.1.5 ... encore la projection stéréographique ! . . . 33
2.1.6 Transformations . . . 33
2.1.7 Dynamique holomorphe . . . 35
2.2 Polynômes et théorème fondamental de l'algèbre . . . 38
2.2.1 Preuve directe . . . 40
2.2.2 Preuve par homotopie . . . 42
2.2.3 Preuve par relèvement . . . 46
2.3 Conclusion : la conception mathématique de la vérité . . . 48

3 Analyse de Fourier : l'ubiquité d'une représentation **49** ***Séances 3 et 4***
3.1 Un problème pratique : la compression d'images . . . 50
3.1.1 Préliminaires : l'encodage naïf des images . . . 50
3.1.2 Compression . . . 50
3.1.3 Le format JPEG, un simple changement de repère . . . 53
3.2 Une base orthonormale pertinente . . . 58
3.2.1 Le produit scalaire, mesure de corrélation . . . 58
3.2.2 Transformée de Fourier continue . . . 64
3.3 Une base adaptée à la dérivation . . . 70
3.3.1 Décomposition dans une base d'harmoniques . . . 72
3.3.2 Conclusion . . . 80

Séance 5 **4 Introduction à la géométrie Riemannienne** **81**
4.1 Axiomatiques non-euclidiennes . . . 82
4.2 Une géométrie non-euclidienne . . . 84
4.2.1 Géodésiques du plan euclidien . . . 86
4.2.2 L'exemple du monde sphérique . . . 88
4.3 Métriques locales sur un ouvert de $\mathbb{R}^n$ . . . 90
4.4 Géodésiques du disque de Poincaré . . . 93
4.4.1 Projections et changements de coordonnées . . . 93
4.4.2 Modèles standards de la géométrie hyperbolique . . . 96
4.4.3 Bilan . . . 102
4.5 Intérêt de la géométrie Riemannienne : l'exemple du tore . . . 102
4.6 Conclusion, ouverture vers la géométrie combinatoire . . . 104
4.6.1 Le sixième modèle . . . 106
4.6.2 L'hyperbolicité au sens de Gromov . . . 107

Séance 6 **5 Un espace de formes étonnant : la sphère des triangles** **109**
5.1 Étude rudimentaire d'une population de poissons . . . 112
5.2 Menhirs, Cornouailles et sphère des triangles . . . 116
5.2.1 Un système de coordonnées adaptées . . . 117
5.2.2 Sphère des triangles et distance procustéenne . . . 122
5.2.3 Statistiques sur la sphère . . . 126
5.2.4 Conclusion . . . 128

Séance 7 **6 Un domaine de recherche actuel : l'anatomie computationnelle** **129**
6.1 Au delà des similitudes : les déformations fluides . . . 129
6.1.1 Un point de vue radicalement opposé : le transport optimal . . . 130
6.1.2 Une régularisation qui passe par la géométrie Riemannienne . . . 134
6.1.3 Tir géodésique sur une variété Riemannienne . . . 137
6.1.4 Métriques à noyaux, premières intuitions . . . 140
6.1.5 Des cométriques à noyaux aux déformations fluides . . . 144
6.1.6 Décomposition de la variabilité anatomique : l'algorithme complet . . . 152
6.1.7 Un véritable programme informatique . . . 154
6.2 Applications en imagerie médicale . . . 166
6.3 Le travail du mathématicien appliqué . . . 172

II Fondements des mathématiques : de la logique aux distributions **175**

Séances 8 et 9 **7 Preuves formelles, axiomatiques et théorie des ensembles** **177**
Revoir l'intro.
7.1 Formules logiques . . . 178
7.2 Axiomatiques et vérités . . . 179
7.2.1 Démonstrations formelles . . . 179
7.2.2 Exemple fondamental : La théorie des ensembles . . . 181
7.2.3 Vérité sémantique et théorème de complétude de Gödel . . . 185
7.3 In-décidabilité, choix d'un système d'axiomes . . . 185
7.3.1 Ni démontrable, ni faux : le paradoxe de l'indécidabilité . . . 186
7.3.2 Les théorèmes d'incomplétude de Gödel . . . 186
7.3.3 Tout est ensemble (?) . . . 190

8 Construction classique des ensembles de nombres : $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ et $\mathbb{R}$ **191** ***Séances 10 et 11***
8.1 Les entiers naturels : successeur et récurrence . . . 191
8.1.1 Axiomes de Peano . . . 192
8.1.2 Construction de Von Neumann . . . 193
8.2 Entiers relatifs et nombres rationnels : premières structures algébriques . . . 196
8.2.1 Les entiers relatifs facilitent la mise en équations . . . 196
8.2.2 Les nombres rationnels permettent de résoudre les problèmes linéaires . . . 198
8.3 Les nombres réels ou la puissance du continu . . . 199
8.3.1 Problème des valeurs intermédiaires . . . 199
8.3.2 Le continu existe-t-il ? . . . 202
8.3.3 Les coupures de Dedekind . . . 203

9 Histoire du calcul différentiel **207** ***Séance 12***
9.1 Indivisibles de Cavalieri . . . 208 ***À réécrire.***
9.2 Pesée d'Archimède . . . 209
9.3 Le calcul différentiel . . . 210
9.3.1 De l'attraction gravitationnelle aux lois de Kepler . . . 210
9.3.2 Abstraction par le calcul infinitésimal de Leibniz . . . 212
9.3.3 Limites de Weierstrass et sommes de Riemann . . . 215
9.3.4 Quadrature analytique de la parabole . . . 219

10 Dimension infinie, dualité et méthode des éléments finis **221** ***Séance 13***
10.1 Compacité en grande dimension . . . 222
10.2 Les fonctions forment un espace de dimension infinie . . . 224
10.3 Les distributions de Schwartz, fonctions généralisées . . . 226
10.4 Comment dériver ce qui n'est même pas continu ? . . . 230
10.5 La méthode des éléments finis . . . 233
10.5.1 Des équations différentielles pour modéliser le monde physique . . . 233
10.5.2 Résolution numérique d'équations aux dérivées partielles . . . 236

Appendices **241**

A Arithmétique **243**

B Éléments de topologie : continuité, limites et points fixes **245**
B.1 La notion de limite . . . 245
B.2 Continuité . . . 245
B.2.1 Théorème des valeurs intermédiaires . . . 245
B.3 Compacité . . . 245
B.4 Deux illustrations : les théorèmes de point fixe . . . 245
B.4.1 Le théorème de Brouwer . . . 245
B.4.2 Le théorème de Cauchy-Lipschitz . . . 245

Chapitre 1

Introduction

Séance 1

Le plan a changé !

Que sont les mathématiques ?

À en croire certains manuels scolaires, une simple collection de *faits*, de recettes sûres qu'il convient de mémoriser afin de "se préparer pour l'année prochaine". À l'opposé, certains auteurs soutiennent qu'il s'agit d'un art pur, peut-être le plus élevé d'entre tous – on lira avec plaisir l'excellent "A Mathematician's Lament" de Paul Lockhart... Sans tomber dans ces deux extrêmes, je tâcherai ici de vous faire découvrir la richesse, la profondeur d'une science bien trop méconnue du public.

Mathématiques et arts graphiques Personne ne penserait aujourd'hui réduire le *dessin* aux simples représentations sans âme du dessin technique : les musées, les bandes-dessinées, tout nous informe que « dessiner », c'est beaucoup plus que « décalquer » – quelques exemples sont donnés Figure 1.3.

Malheureusement, les mathématiques n'ont jamais bénéficié d'une telle reconnaissance. Quel élève de collège a déjà rencontré une *démonstration* du fameux théorème de Pythagore ? À ses yeux, « comprendre » est devenu synonyme d'« apprendre » ; « démontrer », c'est « remplir les trous dans une formule du cours ». Plaisir, élégance et créativité se font de plus en plus rares dans nos salles de classe : on ne s'étonne plus du dégoût pour une matière qui paraît seulement destinée à trier les jeunes pousses. Mais à vous qui côtoyez chaque jour de futurs mathématiciens, je voudrais faire connaître la vérité : non, nous ne passons pas les journées à "calculer" ; non, la "rigueur" ne nous a pas rendu obtus et carrés ; et oui, nous sommes aussi des rêveurs.

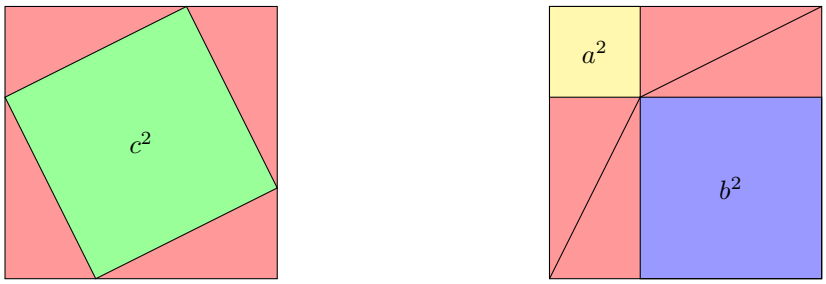
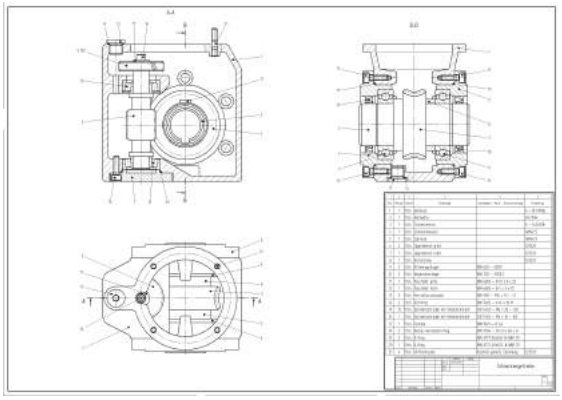


FIGURE 1.1 – Une démonstration – parmi tant d'autres ! – du théorème de Pythagore.

(a) Relié pour enfants.



(b) Dessin technique.

FIGURE 1.2 – Le dessin n'est pas qu'un moyen de brider la créativité des enfants, ou un outil de production industrielle...

(a) *Melencolia I* de Dürer. Cent interprétations pour une seule gravure.

(b) *La grande vague* d'Hokusai : la nature en majesté.

(c) *Maus* d'Art Spiegelmann. En délaissant le réalisme pour un langage graphique épuré mais *puissant*, l'auteur fait passer des idées fortes de manière simple, poignante. Ici, la dualité juif-souris vs nazi-chat.

FIGURE 1.3 – C'est aussi – avant tout ! – un art, un moyen de communication qui donne du sens au monde en le "simplifiant". Il en va de même des mathématiques, et ce cours est là pour vous le faire découvrir. Au delà des "calculs", un monde d'analogies porteuses de sens.

Le jeu de taquin : les étapes du raisonnement mathématique

Avant d'énoncer de grandes généralités creuses sur le plan du cours, il me faut vous faire entrevoir ce dont je veux parler, trouver un exemple. Quoi de mieux qu'un petit jeu symbolique ? Ces passe-temps de l'esprit ont de nombreux points communs avec le travail mathématique, et on ne compte plus les mathématiciens distingués dans leurs jeunesses pour leurs parties d'échecs ou leurs versions latines...

Règles et problème de Loyd

Dans cette séance d'introduction, nous nous concentrerons sur l'un des puzzles mathématiques les plus populaires de la Belle Époque : le jeu de taquin, inventé aux États-Unis dans les années 1870. Ancêtre en deux dimensions du célèbre Rubik's Cube, il s'agit d'un jeu de *permutation* : partant d'une répartition aléatoire de 15 tuiles sur un plateau carré 4x4, le joueur tâchera, par simples glissements, de retrouver l'image cachée dans le puzzle – voir figure 1.4.

Posée par l'inventeur Sam Loyd avec une récompense de 1000$ à la clé, la question que nous chercherons à résoudre est alors la suivante :

« Existe-t-il des configurations de départ qui ne permettent pas au joueur de résoudre le puzzle sans tricher – i.e. sans soulever de tuile ? »

Inversement (les deux questions sont équivalentes) :

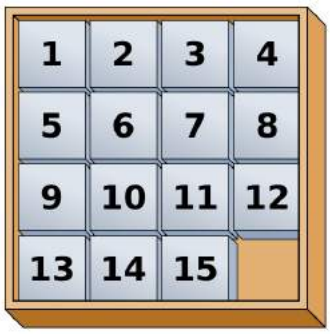
« Tout réarrangement des tuiles est-il accessible par simples glissements à partir de la configuration de départ "ordonnée" ? »

Formalisation des règles

Afin d'apporter une réponse au problème de Loyd, il convient d'abord de formaliser correctement notre problème, en réduisant le jeu de taquin à une idéalisation débarrassée du gras inutile que sont les notions d'image ou de glissement.

Numération Plutôt que de parler de tête, mains, bouche du petit bonhomme de la figure 1.4.a, on représentera notre jeu par un tableau de tuiles numérotées suivant un ordre arbitraire. Une *configuration* sera alors la donnée d'un remplissage du tableau de 16 cases ; on en dénombre exactement $16! = 16 \times 15 \times 14 \times \cdots \simeq 2 \times 10^{13}$ – il y a 16 choix dans $\{1, 2, \ldots, 15, \square\}$ pour la

(a) Jeu de taquin en bois, tiré de `www.jeuxpicards.org/`.



(b) Schéma en 3D isométrique, tiré de Wikipédia.

4	3	2	1
5	6	7	8
12	11	10	9
13	14	15	

(c) La même situation de départ, avec une numérotation adaptée à la résolution du problème de Loyd.

FIGURE 1.4 – Présentation du jeu de taquin.

Le plan a changé !

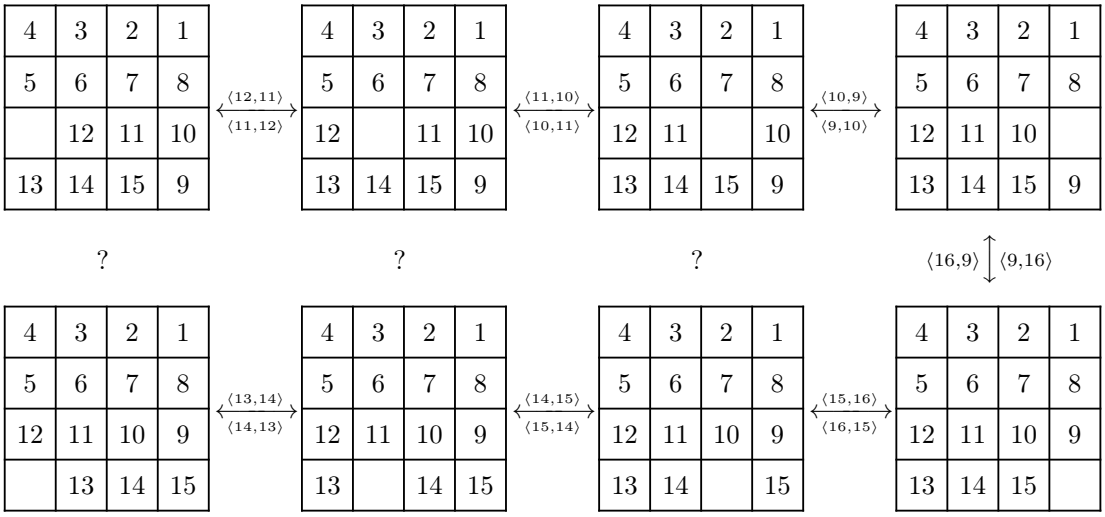


FIGURE 1.5 – Règles du jeu de taquin : les règles $\langle i, j\rangle$ qui permettent de faire glisser la case vide de la i$^{\text{e}}$ à la j$^{\text{e}}$ position sont ici illustrées à partir de la configuration de départ, en bas à droite. Notez la non-commutativité du procédé : aucune règle ne permet de passer directement des trois configurations en haut à gauche à celles du bas.

première case, puis 15 pour la deuxième (qui ne peut prendre la même valeur que la première), 14 pour la troisième, etc. On dit que l'*ensemble* des configurations C a 16! éléments, et pour toute configuration c de C et tout indice i, « $c(i)$ » dénotera la valeur de la tuile associée par c à la position i.

Une *configuration* étant modélisée par une permutation des tuiles du tableau, les règles sont alors données par un jeu de relations $\langle i, j\rangle$ définies sur notre ensemble C : celles-ci encodent le glissement de la tuile vide « □ » de la case i à la case *voisine* j.

Formellement, on définit les ensembles de déplacements horizontaux et verticaux

$$\begin{aligned}
D_{\text{hor}} = \{ \quad & (4,3), \quad (3,2), \quad (2,1), && (1.1)\\
& (5,6), \quad (6,7), \quad (7,8), && (1.2)\\
& (12,11), (11,10), \quad (10,9), && (1.3)\\
& (13,14), (14,15), (15,16) \ \}, && (1.4)\\
D_{\text{ver}} = \{ \quad & (4,5), \quad (5,12), (12,13), && (1.5)\\
& (3,6), \quad (6,11), (11,14), && (1.6)\\
& (2,7), \quad (7,10), (10,15), && (1.7)\\
& (1,8), \quad (8,9), \quad (9,16) \ \}. && (1.8)
\end{aligned}$$

qu'il convient de symétriser – la case vide glisse aussi bien en avant qu'en arrière –

$$\begin{aligned}
D^{\text{sym}}_{\text{hor}} &= \{\, (i,j) \mid (i,j) \in D_{\text{hor}} \text{ ou } (j,i) \in D_{\text{hor}} \,\}, && (1.9)\\
D^{\text{sym}}_{\text{ver}} &= \{\, (i,j) \mid (i,j) \in D_{\text{ver}} \text{ ou } (j,i) \in D_{\text{ver}} \,\}. && (1.10)
\end{aligned}$$

L'ensemble des 48 *déplacements admissibles* est alors la réunion :

$$D = D^{\text{sym}}_{\text{hor}} \cup D^{\text{sym}}_{\text{ver}}. \tag{1.11}$$

La construction verbeuse ci-dessus nous a permis d'écrire noir sur blanc quels déplacements sont autorisés ou interdits par les règles du taquin. Pour tout déplacement admissible (i, j) de D, on définit alors l'application de glissement

$$g_{i\to j} \ : \{\, c \in C \mid c(i) = \square \,\} \to \{\, c \in C \mid c(j) = \square \,\} \qquad (1.12)$$
$$c \mapsto \left(g_{i\to j}(c) \ : \ k \mapsto \begin{cases} \square & \text{si } k = j \\ c(j) & \text{si } k = i \\ c(k) & \text{sinon} \end{cases} \right)$$

qui associe simplement à une configuration c (donnée avec la case vide en i$^{\text{e}}$ position) la configuration $g_{i\to j}(c)$ obtenue en faisant glisser le carré vide sur la j$^{\text{e}}$ case – un véritable joueur ferait plutôt glisser la tuile de la j$^{\text{e}}$ à la i$^{\text{e}}$ case, mais l'écrire ainsi rendrait le suivi de l'espace libre assez fastidieux.

On peut alors définir la relation $\langle i, j \rangle$ par

$$c \xrightarrow{\langle i,j \rangle} d \quad \Leftrightarrow \quad d = g_{i\to j}(c). \qquad (1.13)$$

et on dira que deux configurations c et d sont *joignables*, ce que l'on note « $c \leftrightarrow d$ », s'il existe un chemin de déplacements admissibles $(i_1, j_1), \ldots, (i_n, j_n)$ tel que

$$c \xrightarrow{\langle i_1,j_1 \rangle} g_{i_1\to j_1}(c) \xrightarrow{\langle i_2,j_2 \rangle} \cdots\cdots \xrightarrow{\langle i_n,j_n \rangle} d. \qquad (1.14)$$

Problème de Loyd Le fastidieux travail du paragraphe précédent avait pour but de montrer qu'il est possible d'*encoder* le problème de Loyd en un énoncé mathématique bien formulé :

« Toute configuration est-elle *joignable* à la configuration ordonnée ? »

En pratique, cette triviale étape d'encodage est souvent expédiée en quelques lignes : le véritable raisonnement mathématique n'est pas là, mais dans les pages qui suivent.

Choix d'une représentation Le premier pas, peut-être le plus important, est de passer d'un simple *encodage* à une véritable *représentation* de haut niveau adaptée au problème. Raisonner sur les glissements $g_{i\to j}$ est en effet difficile : à cause de la condition sur la case vide, impossible par exemple de composer deux glissements quelconques, comme $(1, 2)$ et $(3, 4)$... Tenter d'étudier directement un jeu de relations aussi hétéroclite que celui des $\langle i, j \rangle$, c'est la garantie de s'empêtrer dans les cas particuliers sans faire émerger d'idée forte.

Équivalence La première avancée sera de décomposer le jeu de déplacements D en une partie *triviale*, composée des mouvements de la case vide le long du serpentin $(1, \ldots, 16)$, et une partie *non-triviale*, qui comprend les mouvements verticaux qui ne sont pas de la forme $(i, i \pm 1)$.

On dira alors que deux configurations c et d sont *équivalentes*, ce que l'on note « $c \sim d$ », si c et d sont joignables par un chemin de déplacements *triviaux*. Autrement dit, deux configurations sont équivalentes si et seulement si leurs tuiles non-vides sont rangées dans le même ordre par rapport à la numérotation des cases – voir figure 1.6.

Passage au quotient La relation d'*équivalence* est simple à visualiser : chaque configuration c est équivalente à exactement 15 autres configurations qui correspondent aux différentes positions de la case vide le long du serpentin. Plutôt que de continuer à s'encombrer de la donnée de cette position – que l'on peut faire varier à loisir via les déplacements *triviaux* de la forme $(i, i \pm 1)$ –, on

Le plan a changé !

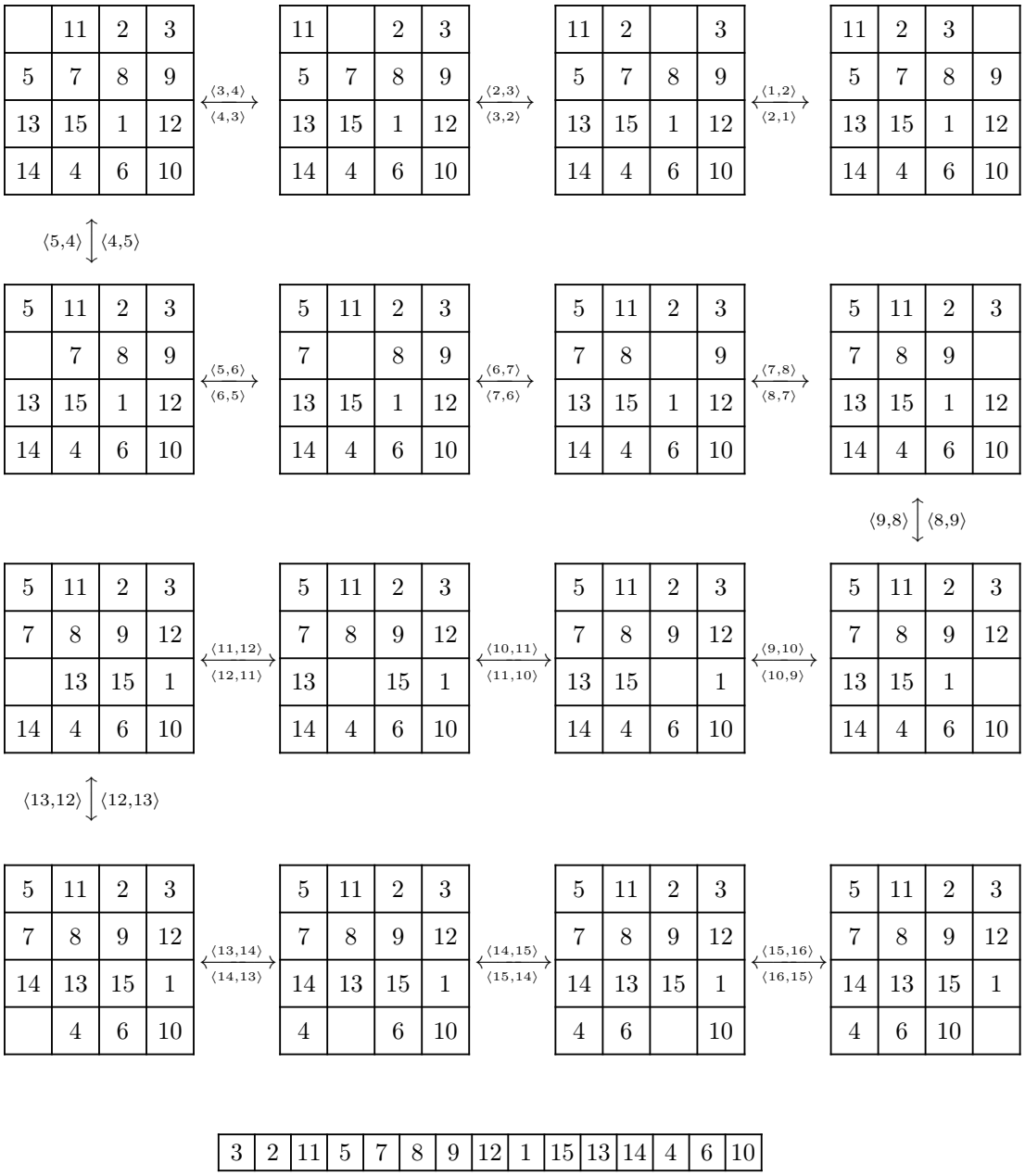


FIGURE 1.6 – Exemple d'utilisation des relations $\langle i, i \pm 1 \rangle$. Toutes les configurations affichées ici sont identifiées à la permutation $[3, 2, 11, 5, 7, 8, 9, 12, 1, 15, 13, 14, 4, 6, 10]$.

va donc s'en débarrasser en travaillant directement sur l'information résiduelle, ou « *quotient* » : la donnée $\sigma(c)$ de l'arrangement ordonné des 15 tuile non vides, tel que lu le long du serpentin.

Formellement le passage au quotient s'apparente à une factorisation :

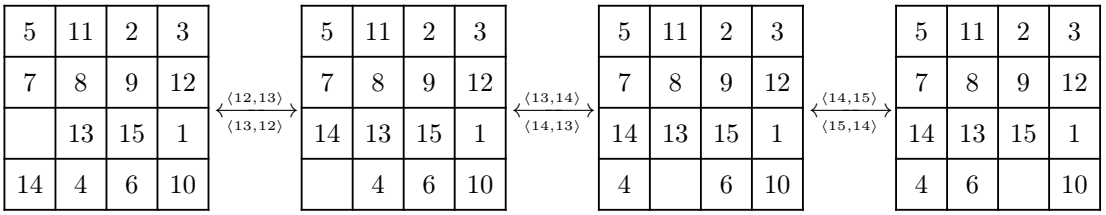
$$\begin{aligned} F \ : C &\to \ S_{15} \times \{\,1,\dots,16\,\} \\ c &\mapsto (\,\sigma(c), c^{-1}(\square)\,) \end{aligned} \tag{1.15}$$

où
$$C = \left\{ \text{ arrangements de } \{\,1,\dots,15,\square\,\} \text{ sur } \{\,1,\dots,16\,\} \right\}, \tag{1.16}$$
$$S_{15} = \left\{ \text{ arrangements de } \{\,1,\dots,15\} \text{ sur } \{\,1,\dots,15\,\} \right\}, \tag{1.17}$$
$$c^{-1}(\square) \quad \text{désigne la position de la case vide dans } c. \tag{1.18}$$

Comme F est bijective – à toute configuration correspond une unique factorisation, et réciproquement –, on peut dire qu'elle est une *représentation* sur laquelle les relations *triviales* s'expriment simplement : pour toute configuration c identifiée à son image $(\sigma_c, \square_c)$ par F, pour tout déplacement trivial $(\square_c, \square_c \pm 1)$, on a

$$g_{\square_c \to \square_c \pm 1}(\sigma_c, \square_c) = (\sigma_c, \square_c \pm 1)\,; \tag{1.19}$$

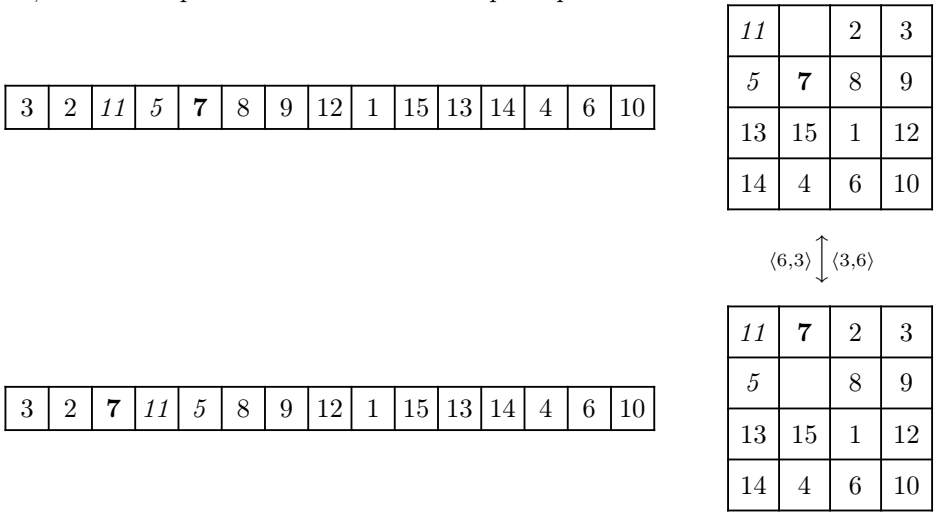
autrement dit, un déplacement trivial permet de modifier $\square_c$ à loisir, sans toucher à σ_c.



(σ, 12) (σ, 13) (σ, 14) (σ, 15)

Action triviale des relations $(i, i \pm 1)$ dans la factorisation donnée équation (1.15).
On a ici $\sigma = [3, 2, 11, 5, 7, 8, 9, 12, 1, 15, 13, 14, 4, 6, 10]$.

Action des déplacements non-triviaux Les 2x9 déplacements restants ont une action plus intéressante : non-contents de faire varier la position $\square_c$ d'un nombre pair, ils agissent aussi sur σ_c par "sauts de moutons". La règle $\langle 3, 6\rangle$ permet par exemple de faire passer la 5e tuile en 3e position, en sautant par-dessus les deux tuiles qui la précédent :



On peut ainsi établir le catalogue exhaustif des actions sur la première composante σ, dont les inverses se déduisent immédiatement et qui sont *concaténables* à envie – puisque les règles triviales permettent de faire bouger la case vide où on le souhaite sur le plateau :

$\langle 3,6\rangle$: avance la 5^{e} tuile en 3^{e} position, en sautant par dessus les 2 tuiles qui la précèdent,
$\langle 2,7\rangle$: avance la 6^{e} tuile en 2^{e} position, en sautant par dessus les 4 tuiles qui la précèdent,
$\langle 1,8\rangle$: avance la 7^{e} tuile en 1^{re} position, en sautant par dessus les 6 tuiles qui la précèdent,

$\langle 7,10\rangle$: avance la 9^{e} tuile en 7^{e} position, en sautant par dessus les 2 tuiles qui la précèdent,
$\langle 6,11\rangle$: avance la 10^{e} tuile en 6^{e} position, en sautant par dessus les 4 tuiles qui la précèdent,
$\langle 5,12\rangle$: avance la 11^{e} tuile en 5^{e} position, en sautant par dessus les 6 tuiles qui la précèdent,

$\langle 11,14\rangle$: avance la 13^{e} tuile en 11^{e} position, en sautant par dessus les 2 tuiles qui la précèdent,
$\langle 10,15\rangle$: avance la 14^{e} tuile en 10^{e} position, en sautant par dessus les 4 tuiles qui la précèdent,
$\langle 9,16\rangle$: avance la 15^{e} tuile en 9^{e} position, en sautant par dessus les 6 tuiles qui la précèdent.

Signature d'une permutation

Dans la représentation F le problème de Loyd se reformule simplement :

« Tout couple $(\sigma_c, \square_c)$ est-il *joignable* à la configuration ordonnée donnée par le couple $([1,2,\ldots,15],16)$? »

Les règles *triviales* permettant de modifier la deuxième coordonnée de nos couples à loisir, cette question se restreint au seul problème sur la première coordonnée σ :

« Toute permutation $\sigma \in S_{15}$ est-elle *joignable* à la permutation ordonnée $[1,2,\ldots,15]$ via l'action des règles non-triviales ? »

On a donc ramené le problème du *taquin* à un simple problème de *tri* de listes ordonnées, la file de "cartes" numérotées de 1 à 15 ayant pris la place du complexe plateau aux 16 tuiles coulissantes.

Réponse à la question de Loyd Surtout, l'écriture des règles non-triviales a permis de faire émerger une propriété remarquable, qui ne sautait pas aux yeux équation (1.12) : le nombre de tuiles "sautées" par l'application d'une règle du jeu de taquin est toujours *pair*. Autrement dit, s'il est possible de dépasser 2, 4, ou 6 tuiles, aucune règle ne permet de permuter deux tuiles consécutives. Ce constat est au cœur du théorème suivant :

Théorème 1.1. *(Johnson, 1879) La réponse au problème de Loyd est* négative *: il est impossible de joindre la configuration ordonnée à la configuration "impaire" où les tuiles 1 et 2 ont été échangées.*

Si l'intuition est correcte, il reste évidemment à en tirer une démonstration : après tout, comment assurer qu'une combinaison astucieuse des sauts-de-moutons ne nous amènera pas par des voies détournées à une configuration "impaire" ? L'outil de preuve le plus élégant sera ici la notion d'*invariant*.

1	**2**	3	4	5	6	7	8	9	10	11	12	13	14	15

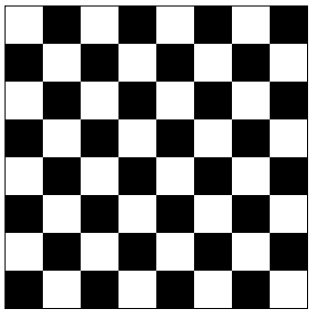
4	3	**2**	*1*
5	6	7	8
12	11	10	9
13	14	15	

↕ Impossible

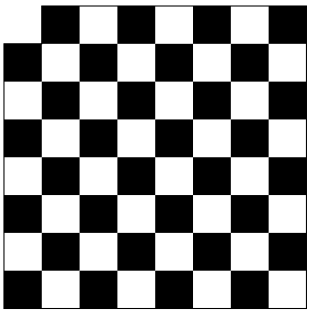
2	*1*	3	4	5	6	7	8	9	10	11	12	13	14	15

4	3	*1*	**2**
5	6	7	8
12	11	10	9
13	14	15	

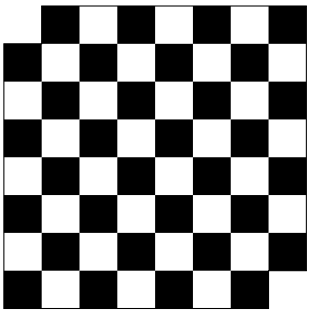
Dominos et damiers En théorie des jeux, un *invariant* est une quantité – calculable à partir de toute configuration du plateau – qui reste constante quelque soit le mouvement choisi par le joueur. Considérons par exemple le problème de *pavage* du damier 8x8 par des dominos. Il n'est pas difficile de trouver une solution dans le cas où le damier est complet. Par contre, si un des coins est rogné, l'impossibilité de résoudre le puzzle saute aux yeux : un jeu de dominos ne peut recouvrir un nombre impair de cases ! Mais alors, quid du cas où les deux coins blancs sont rognés ?



(a) Damier 8x8.



(b) Avec un coin blanc tronqué.



(c) Avec les deux coins blancs tronqués.

À forces d'essais, on peut se convaincre que ce problème est lui aussi insoluble ; mais il est difficile de s'en assurer : un raisonnement par récurrence/induction ad hoc est voué à s'enliser dans les cas particuliers. Non, *la* bonne manière de procéder est de considérer la quantité :

$$I(p) = \text{nbre. cases noires visibles} - \text{nbre. cases blanches visibles}. \tag{1.20}$$

On sait que pour un damier pavé, $I(p) = 0 - 0 = 0$, et on calcule sans peine que $I(\text{« deux coins blancs tronqués »}) = 32 - 30 = 2$. Or I est un invariant du jeu de pavage, car poser un domino sur le damier cache toujours exactement une case blanche et une case noire : c'est donc que le pavage du damier tronqué par des dominos est impossible ; Cqfd.

Le plan a changé !

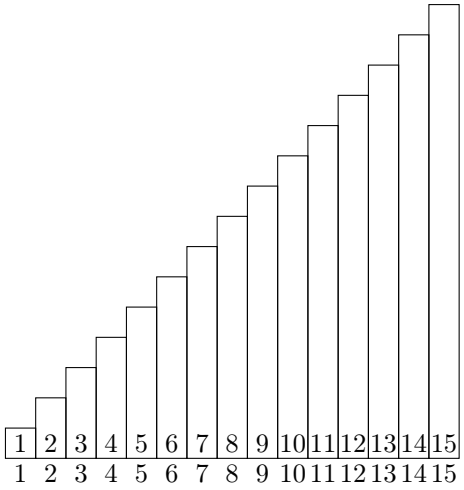
Nombre d'inversions Trouver un invariant au problème de Loyd est autrement plus difficile que d'effectuer une simple soustraction. Il s'agit d'obtenir une formule suffisamment *lâche* pour être invariante par toutes les règles du jeu de taquin, et suffisamment *fine* pour discriminer les deux configurations du théorème 1.1.

La quantité qui va nous tirer d'affaire est la suivante :

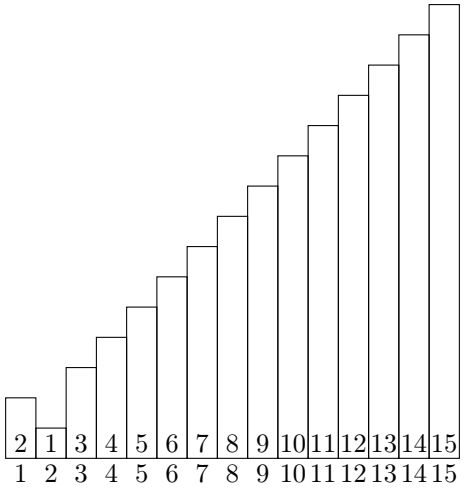
Définition 1.1 (Nombre d'inversions)**.** Si $\sigma \in S_{15}$, on définit son nombre d'inversions $I(\sigma)$ comme le nombre de couples $(i, j) \in \{1, \dots, 15\} \times \{1, \dots, 15\}$ tels que :

$$i < j \qquad \text{et} \qquad \sigma(i) > \sigma(j). \tag{1.21}$$

On calcule aisément les nombres d'inversions associés aux permutations du théorème 1.1 :



(a) Permutation ordonnée : aucune inversion.



(b) Ici par contre, une inversion : $(1, 2)$.

Reste alors à démontrer le lemme suivant :

Lemme 1.1. *La parité du nombre d'inversions de σ est invariante par application des règles du jeu de taquin.*

Preuve Soit $\sigma \in S_{15}$ une permutation de $\{1, \dots, 15\}$, et considérons l'opération de "saute-mouton", ou *transposition*, qui échange la k$^{\text{e}}$ tuile avec sa voisine pour la mettre en position $k - 1$. On note τ la permutation obtenue à partir de σ.

Pour tout couple $(i, j) \in \{1, \dots, 15\} \times \{1, \dots, 15\}$ avec i et j tous deux différents de k et $k - 1$,

$$\begin{aligned}
(i, j) \text{ est une inversion pour } \sigma &\Leftrightarrow (i < j) \text{ et } \sigma(i) < \sigma(j) && (1.22)\\
&\Leftrightarrow (i < j) \text{ et } \tau(i) < \tau(j) && (1.23)\\
&\Leftrightarrow (i, j) \text{ est une inversion pour } \tau && (1.24)
\end{aligned}$$

puisque $\tau(i) = \sigma(i)$ et $\tau(j) = \sigma(j)$. Par contre, si $(k - 1, k)$ était une inversion pour σ, alors elle ne l'est plus pour τ – qui a remis $\sigma(k - 1)$ et $\sigma(k)$ dans le bon ordre –, et réciproquement : si $(k - 1, k)$ n'était pas une inversion pour σ, alors elle le devient pour τ – qui renverse $\sigma(k - 1)$ et

$\sigma(k)$. Enfin, pour tout indice $i \notin \{k-1, k\}$, on a

$$(i,k) \text{ est une inversion pour } \sigma \Leftrightarrow (i,k-1) \text{ est une inversion pour } \tau \tag{1.25}$$
$$(k,i) \text{ est une inversion pour } \sigma \Leftrightarrow (k-1,i) \text{ est une inversion pour } \tau \tag{1.26}$$
$$(i,k-1) \text{ est une inversion pour } \sigma \Leftrightarrow (i,k) \text{ est une inversion pour } \tau \tag{1.27}$$
$$(k-1,i) \text{ est une inversion pour } \sigma \Leftrightarrow (k,i) \text{ est une inversion pour } \tau \tag{1.28}$$
$$\tag{1.29}$$

Si on récapitule, l'ensemble des inversions de τ est donc constitué :

— des couples (i,j) qui n'intersectent pas $\{k-1,k\}$ et qui sont des inversions de σ ;

— des couples (i,k) (resp. (k,i)), pour $i \neq k-1$ et $(i,k-1)$ (resp. $(k-1,i)$) inversion de σ ;

— des couples $(i,k-1)$ (resp. $(k-1,i)$), pour $i \neq k$ et (i,k) (resp. (k,i)) inversion de σ ;

— du couple $(k-1,k)$ si $(k-1,k)$ n'est *pas* une inversion de σ.

On a donc $I(\tau) = I(\sigma) \mp 1$, selon que $(k-1,k)$ est ou non une inversion pour σ. Dans tous les cas, la parité de $I(\tau)$ et de $I(\sigma)$ n'est pas la même. Or on a vu à la section précédente que les règles du taquin pouvaient toujours être assimilées à un nombre *pair* de transpositions entre tuiles voisines :

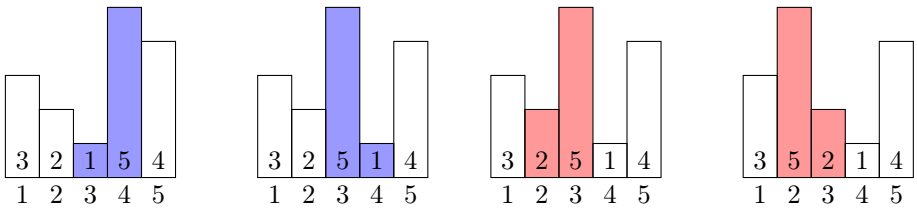
— 0 si la règle est triviale ;

— 2 s'il s'agit de $\langle 3,6\rangle, \langle 7,10\rangle, \langle 11,14\rangle$ ou de leurs inverses ;

— 4 s'il s'agit de $\langle 2,7\rangle, \langle 6,11\rangle, \langle 10,15\rangle$ ou de leurs inverses ;

— 6 s'il s'agit de $\langle 1,8\rangle, \langle 5,12\rangle, \langle 9,16\rangle$ ou de leurs inverses.

Changer de parité $2n$ fois, c'est ne pas changer de parité du tout : la parité du nombre d'inversions est donc un invariant des règles du taquin ; Cqfd. Le théorème 1.1 est donc démontré.



(a) 4 inversions :
— $(1,2)$,
— $(1,\mathbf{3})$, $(2,\mathbf{3})$,
— $(\mathbf{4},5)$.

(b) 5 inversions :
— $(1,2)$,
— $(1,\mathbf{4})$, $(2,\mathbf{4})$,
— $(\mathbf{3},5)$,
— $(\mathbf{3},\mathbf{4})$.

i.e.
— $(1,4)$,
— $(1,\mathbf{2})$, $(\mathbf{2},4)$
— $(\mathbf{3},4)$, $(\mathbf{3},5)$.

(c) 6 inversions :
— $(1,4)$,
— $(1,\mathbf{3})$, $(\mathbf{3},4)$,
— $(\mathbf{2},4)$, $(\mathbf{2},5)$,
— $(\mathbf{2},\mathbf{3})$.

FIGURE 1.9 – Un saut de mouton change la parité du nombre d'inversions, un double saut la conserve : de (a) à (b), $k=4$; de (b) = (c), $k=3$.

Le plan a changé !

Pour aller plus loin : plan du cours

Abstraction Cette résolution du problème de Loyd aura été l'occasion de mettre en avant deux aspects fondamentaux d'une bonne démonstration mathématique : L'*élégance* du raisonnement, d'abord, qui permet d'apporter une solution limpide à un problème confus ; trouver un invariant approprié au jeu de taquin nous aura permis de répondre au problème sans avoir à explorer l'arbre de toutes les parties possibles. Mais surtout, soulignons le rôle essentiel joué par la *représentation* des données, qui aura clarifié des règles de jeu a priori peu propices aux raisonnements.

Les mathématiques forment une science qui donne du *sens* aux situations pratiques en identifiant en elles des *structures* simples, compréhensibles : du jeu de taquin, un simple problème de tri. "Abstraire" un problème, c'est l'extraire d'un carcan de détails pratiques pour en isoler l'armature intrinsèque. Étudier ces structures pour elles-mêmes est alors l'objet des mathématiques *fondamentales*, tandis que les mathématiciens *appliqués* tirent de l'unité structurelle entre des domaines pratiques variés de nombreux résultats inattendus, des raisonnements par analogie surprenants.

Objectifs du cours Comme illustré Figure 1.10, les mathématiques sont aujourd'hui subdivisées en de nombreux domaines plus ou moins appliqués, plus ou moins "formels"... Chaque branche correspond en fait à l'étude d'une notion, d'une structure "abstraite" et des questions qui en découlent naturellement. La topologie s'intéressera par exemple à la notion de *continuité* et la logique, à celle de *preuve formelle*.

Proposer un cours de "culture mathématique" est bien ambitieux : depuis le début du XX^{e} siècle, le nombre de mathématiciens – et avec lui celui des domaines – a explosé. Au vu de la croissance exponentielle du nombre d'articles publiés depuis 1850, on admet d'ailleurs généralement qu'Henri Poincaré – mort en 1912 – fut le dernier homme à pouvoir contribuer effectivement à tous les domaines de son temps. Impossible, donc, d'aborder en un cours l'ensemble des domaines étudiés par les mathématiciens aujourd'hui.

Face à cette difficulté, on peut faire le choix d'une coupe transversale Algèbre–Probas–Géométries, survolant l'équivalent d'un programme de licence de maths : c'est une approche parfaitement légitime, qui permet aux élèves de goûter à toutes les grandes familles de sujets de recherche, d'esprits mathématiques ; mais ce n'est pas celle que nous suivrons ici.

Plutôt que de vous présenter des *sujets*, je voudrais en effet vous proposer une *vision* du monde. Celle des mathématiciens, qui géométrisent les photos souvenirs et probabilisent les réseaux de neurones ; pour qui musique et chaleur ne sont que les deux faces d'une même pièce... On n'y parlera donc pas des éternelles marottes de la vulgarisation mathématique que sont les hôtels de Cantor, les ensembles fractals ou la suite de Fibonacci. Ces sujets sont aux mathématiques ce que les trous noirs sont à la physique : des problèmes intéressants, curieux mais relativement annexes, loin d'être au cœur des préoccupations de la communauté des chercheurs – en dépit d'une large couverture médiatique.

À l'opposé, on optera pour un plan "vertical", qui seul permet de présenter ce qui est pour moi l'essence de cette discipline : la rencontre entre la *rigueur* de la logique formelle, qui définit proprement les objets du discours mathématique, et la capacité d'*abstraction*, qui identifie dans une multitude de phénomènes disparates une même structure sous-jacente.

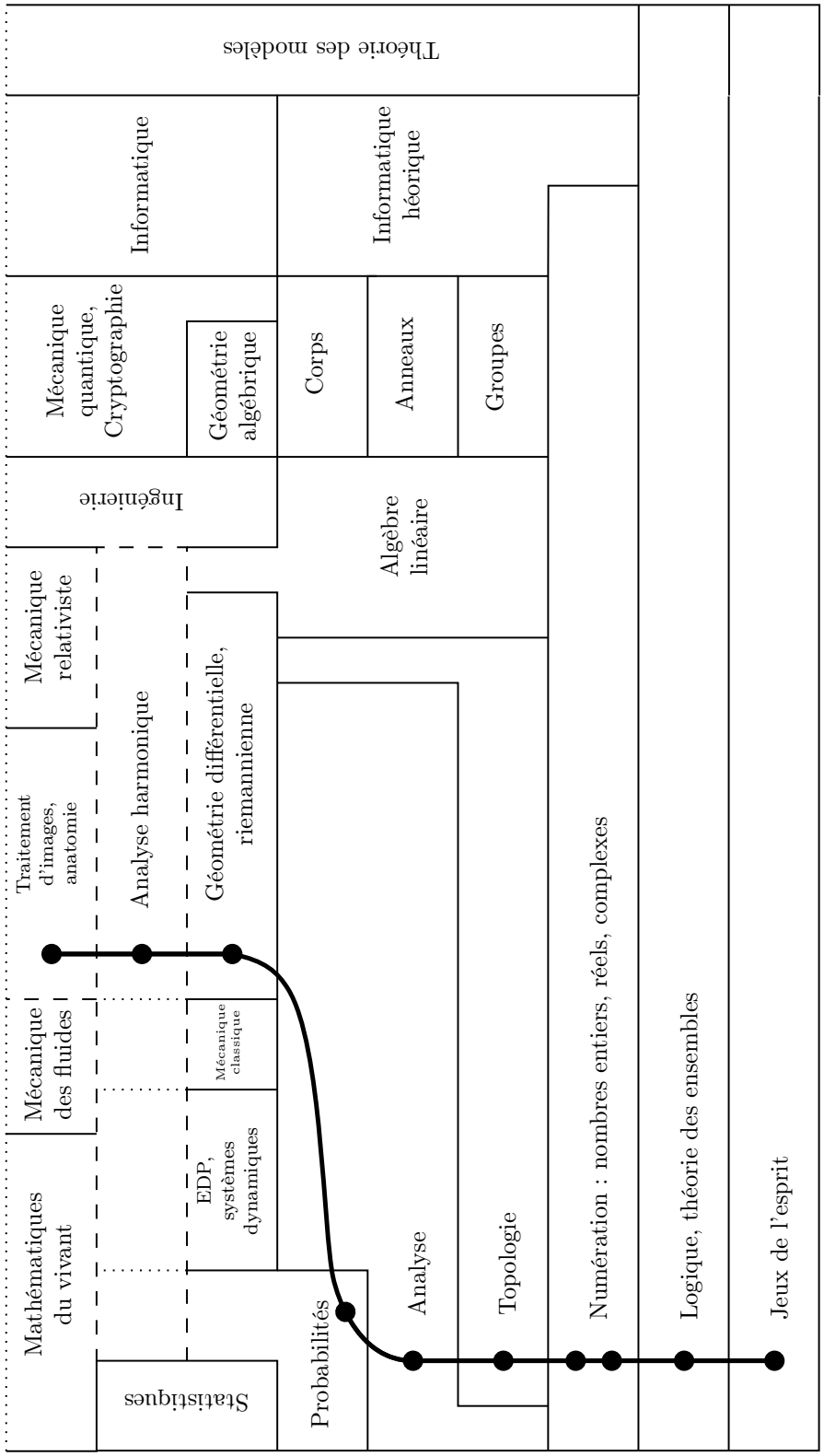


FIGURE 1.10 – Quelques branches des mathématiques : les domaines du haut reposent naturellement sur les outils développés aux échelons les plus fondamentaux. On suivra le chemin en gras, des briques les plus fondamentales aux applications médicales.

Le plan a changé !

"Culture des arts graphiques" En vingt-quatre heures seulement, comment présenterait-on les "arts graphiques" à des élèves qui n'auraient jamais été confrontés qu'à de simples reliés et coloriages ? Un plan chronologique ne me paraît pas souhaitable : que comprendrait à Quentin de La Tour un auditeur qui ne fait pas la différence entre un pastel et un burin ? Et à l'inverse, qui voudrait consacrer un semestre aux seules techniques sans jamais voir de chef-d'œuvre ? L'exercice est difficile et un bon exposé, en ne faisant pas d'impasses graves, perdra forcément en cohérence interne : c'est dommage, mais nécessaire.

À supposer que dix ans d'arts plastiques scolaires aient permis à mes élèves de distinguer un feutre d'un crayon, je commencerais mon cours par une leçon sur la *perspective* et l'*anatomie*. Aborder frontalement ces deux thèmes me semble indispensable pour démystifier la discipline, après des années passées à décalquer sagement les épreuves des manuels. Le dessin reposant sur des bases théoriques solides, s'en affranchir doit relever du *choix*, pas de l'incompétence : présenter ces sujets techniques dès le début du cours permettrait donc de mieux apprécier l'audace des artistes contemporains.

Je consacrerais ensuite une partie aux techniques classiques – gravure, crayonné, fusain,...– en illustrant chaque leçon par les œuvres de quelques grands maîtres. Anciens et modernes seraient mis à contribution, mettant ainsi en lumière leurs différences, la part de créativité propre à chaque individu.

Enfin, je terminerais par une ouverture au monde des arts graphiques contemporains. On passerait une leçon à admirer, décortiquer un chef-d'œuvre de Picasso ou Dali... Mais on n'oublierait pas non plus d'accorder une place importante aux media hybrides que sont le dessin animé et la bande dessinée, nouveaux moyens d'expression qui permettent à de talentueux créateurs de toucher un large public.

C'est ce plan, vous vous en doutez, que je me propose de transposer aux mathématiques.

Fondements des mathématiques Naturellement, on commencera par établir le socle commun à tout l'édifice mathématique : la théorie des preuves, ou *logique formelle*. Devant les paradoxes soulevés par les théorèmes d'incomplétude de Gödel, on s'attachera ensuite à définir rigoureusement, à *construire* les ensembles de nombres : entiers naturels, relatifs, nombres rationnels et réels. Enfin, l'insolubilité de l'équation polynomiale $x^2+1=0$ nous mènera à aborder la notion de nombre *complexe*, qui se révélera étonnamment riche et profonde.

Outils conceptuels modernes Plutôt que d'aborder les questions algébriques, on préférera poursuivre sur la voie de l'analyse et de la géométrie. On commencera donc par étudier les *fonctions* de manière rigoureuse : théorie de la *continuité* – sous une forme étonnamment générale –, suivie d'un aparté historique sur le *calcul différentiel*, qui nous permettra de bien mettre en perspective ce qui est sans doute l'idée la plus féconde de toute l'Histoire des sciences.

Pour terminer cette revue des outils essentiels aux mathématiques modernes, nous parlerons enfin de *probabilités* sous trois angles complémentaires : définition axiomatique, liens avec les statistiques, aide à la modélisation de phénomènes déterministes chaotiques.

Les mathématiques aujourd'hui Les fondations posées, et les outils de base présentés, on pourra alors s'attaquer aux concepts de plus haut niveau. Après une introduction aux géométries non-euclidiennes qui sera l'occasion de vous initier au plaisir de la lecture d'une *très* belle preuve, nous consacrerons une séance à l'*analyse harmonique* sous toutes ses formes : Épicycles de Ptolémée, diffusion de la chaleur, musique et traitement d'images seront unifiées dans une même théorie géométrique dont une application directe se trouve être... Le format d'images JPEG, aujourd'hui présent sur tous les appareils mobiles !

Toutes les cartes en mains, nous terminerons alors notre voyage par la présentation d'un domaine de recherche actuel en mathématiques appliquées : le mien. Au confluent des statistiques, de la mécanique des fluides et de la géométrie riemannienne, l'*anatomie computationnelle* a pour objet l'étude quantitative d'une population de *formes* anatomiques : "courbes de croissance" cardiaques, aide au diagnostic de la maladie d'Alzheimer ou comparaisons de fossiles d'hominidés seront donc au rendez-vous.

Pour la prochaine fois

Nous démarrons sur les chapeaux de roues avec une étude fine de la notion de *preuve*. Pour ne pas être décroché dès les premières minutes, il est indispensable que vous soyez familiarisés avec les notions élémentaires de la logique propositionnelle : connecteurs "et", "non", "ou", quanteurs d'existence. Je ne saurais donc que trop vous conseiller de lire :

— Le très bon cours "Éléments de Mathématiques" de David Delaunay, disponible à l'adresse : `mp.cpgedupuydelome.fr/cours.php?id=4484`.

— La feuille d'exercices attenante : `mp.cpgedupuydelome.fr/pdf/Th%C3%A9orie%20des%20ensembles%20et%20des%20applications.pdf`.

Références

Je tiens à remercier mon prédécesseur, Jeremy Daniel, qui m'a donné l'idée d'aborder le jeu de taquin pour cette leçon inaugurale. Si la démonstration historique est due à W. Johnson (*Notes on the "15" puzzle*, 1879 : `archive.org/details/jstor-2369492`), j'ai privilégié l'approche suivie par Michel Coste dans l'article publié sur le site de vulgarisation *Images des Mathématiques* – en la simplifiant un peu : `images.math.cnrs.fr/Le-jeu-de-taquin-du-cote-de-chez-Galois.html`. Démontrer l'insolubilité du problème de Loyd sans utiliser le vocabulaire des groupes de permutation était une gageure !

Au lecteur curieux, je souhaiterais conseiller la lecture de l'excellent article *Les tonalités musicales vues par un mathématicien* (disponible à l'adresse `culturemath.ens.fr/content/les-tonalit%C3%A9s-musicales-vues-par-un-math%C3%A9maticien`). Écrit par Michel Broué, éminent algébriste et directeur du département de 1986 à 1993, ce petit texte est pour moi une référence de ce que doit être un texte de vulgarisation mathématique : clair, honnête et précis.

Le plan a changé !

Première partie

Nombres complexes, géométrie

Chapitre 2

Le corps des nombres complexes

Séance 2

Il est maintenant temps de clore cette première partie du cours consacrée aux fondements des mathématiques. Au chapitre précédent, nous avons *construit* les ensembles de nombres du collège qui permettent de compter, de calculer, de passer au continu. Reste donc, enfin, à aborder le "sommet" du lycée : le corps des nombres complexes.

Dans un premier temps, nous reverrons en douceur les notions algébriques les plus simples, en reprenant la belle exposition du film *Dimensions*. Les fondamentaux révisés, nous n'enchaînerons alors pas immédiatement sur les applications, qui attendront le chapitre 3 consacré à l'analyse harmonique. Pour terminer dignement cette introduction aux mathématiques modernes, je vous propose plutôt de nous intéresser à la notion de preuve : Au travers de l'étude du théorème fondamental de l'algèbre, nous verrons qu'en mathématiques, la preuve nous apporte bien plus qu'un simple résultat.

Retour sur les nombres complexes vus au lycée

La section qui suit est un copié-collé de `www.dimensions-math.org/Dim_CH5.htm`, supplément détaillé de l'excellent film *Dimensions* par Jos Leys, Étienne Ghys et Aurélien Alvarez. Il va sans dire que nous projetterons les chapitres 5 et 6 en classe, et que je vous invite chaleureusement à le regarder chez vous en entier !

Le présentateur

Les *nombres complexes* constituent l'un des plus beaux chapitres des mathématiques et sont devenus essentiels dans la science. Le chemin de leur découverte n'a pas été aisé et la terminologie employée témoigne de cette difficulté ; on a parlé de nombres impossibles, imaginaires, et le mot "complexe" laisse entendre qu'il n'est pas facile de les comprendre. Heureusement ce n'est plus le cas aujourd'hui : nous pouvons maintenant les présenter d'une manière relativement élémentaire.

Adrien Douady est le présentateur de ces chapitres. Mathématicien exceptionnel, ses contributions sont très variées, et il aimait dire que toutes ses recherches tournaient autour des nombres complexes. Il est en particulier l'un de ceux qui ont fait revivre la théorie des systèmes dynamiques complexes dont nous dirons quelques mots plus loin.

L'une des caractéristiques de cette théorie est qu'elle engendre de très jolis ensembles fractals qu'on peut aujourd'hui représenter grâce aux ordinateurs. Adrien Douady fait partie de ceux qui

ont résolument encouragé la production de ce type d'images, à la fois pour aider le mathématicien dans son travail de recherche et pour populariser les mathématiques dans la société.

On lui doit également un film d'animation mathématique intitulé *La dynamique du lapin* : il aimait baptiser les objets mathématiques de noms étonnants : lapin, avion, shadok etc. Sa disparition récente a profondément attristé la communauté des mathématiciens.

Il est clair que même Adrien Douady ne peut pas expliquer toute la théorie des nombres complexes en deux chapitres de 13 minutes... Ces chapitres ne peuvent pas se substituer au cours d'un professeur, à un livre, ou à une exposition détaillée. Il faut considérer ces chapitres comme des compléments ou des illustrations qui encouragent à en savoir plus ou des rappels pour ceux qui auraient oublié de lointaines leçons passées. Bien sûr, le film cherche avant tout à mettre en évidence le côté géométrique de ces nombres complexes.

Nombres et transformations

Nous avons vu que la droite est de dimension 1 puisqu'on peut se repérer sur une droite avec un nombre, positif à droite de l'origine et négatif à gauche. Les points sont des êtres géométriques et les nombres sont des êtres algébriques. L'idée de penser à des nombres comme des points ou à des points comme des nombres, c'est-à-dire de mélanger l'algèbre et la géométrie, est l'une des idées les plus fécondes des mathématiques. Comme toujours, il n'est pas facile de l'attribuer à un seul homme mais c'est en général à Descartes qu'on attribue cette méthode puissante d'étude de la géométrie par l'algèbre : c'est la naissance de la géométrie algébrique. Si les points d'une droite sont des nombres, on doit pouvoir comprendre géométriquement la signification des opérations élémentaires entre nombres : l'addition et la multiplication. La clé de cette compréhension est dans l'idée de transformation.

Par exemple, soustraire 1 à un nombre x, c'est-à-dire la transformation $x - 1$, est vue géométriquement comme une translation : tous les points sont translatés de 1 vers la gauche. De la même manière, la multiplication par 2 est pensée comme une dilatation.

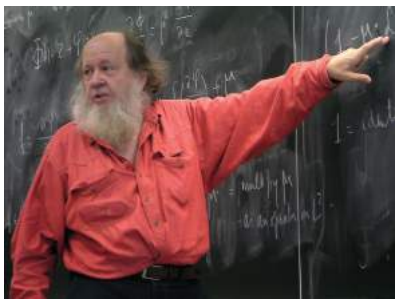
La multiplication par -1 qui envoie chaque point x sur $-x$ est pensée comme une symétrie : chaque point est transformé en son symétrique par rapport à l'origine. La multiplication par -2 est quant à elle la composition des deux opérations précédentes. Multiplier deux nombres revient à composer les transformations qui leur sont associées. Par exemple, la transformation associée à la multiplication par -1 est une symétrie et lorsque l'on effectue cette opération deux fois de suite, on revient au point de départ, si bien que le produit de -1 avec lui-même est $+1$. Le carré de -1 est $+1$.

Le carré de -2 est $+4$ pour la même raison. Il résulte de tout cela que le carré de tout nombre est toujours positif. Il n'y a pas de nombre dont le carré soit égal à -1.

Autrement dit, -1 *n'a pas de racine carrée.*

La racine carrée de -1

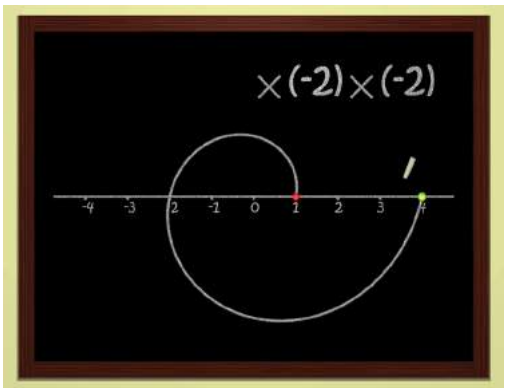
Pendant longtemps, l'impossibilité de trouver une racine carrée pour -1 était un dogme dont on ne pouvait pas discuter. Mais à l'époque de la Renaissance, certains esprits inventifs osèrent rompre le tabou ! Si l'on ose écrire $\sqrt{-1}$, alors on peut aussi écrire des nombres comme par exemple $2 + 3\sqrt{-1}$ et on peut également jouer avec ces nombres de manière formelle, sans trop essayer de comprendre leurs significations. Ces pionniers ont alors constaté de manière en quelque sorte expérimentale que calculer avec ces nombres impossibles ne semblait pas mener à des contradictions si bien que ces nouveaux nombres furent peu à peu acceptés par les mathématiciens, sans de véritables justifications.



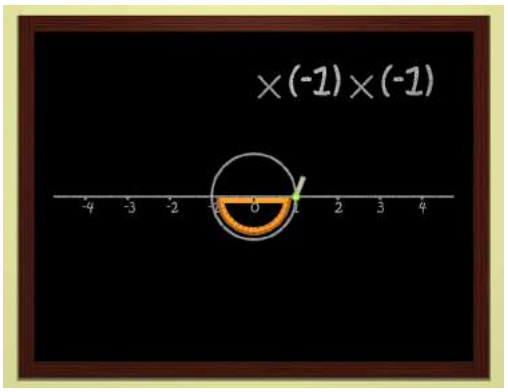
(a) Adrien Douady, par François Tisseyre — Photo prise à l'IHP.

(b) L'un des premiers traités où apparut l'interprétation géométrique des nombres complexes.

FIGURE 2.1 – Le présentateur, et l'œuvre discutée en ce début de chapitre, datant de 1806.

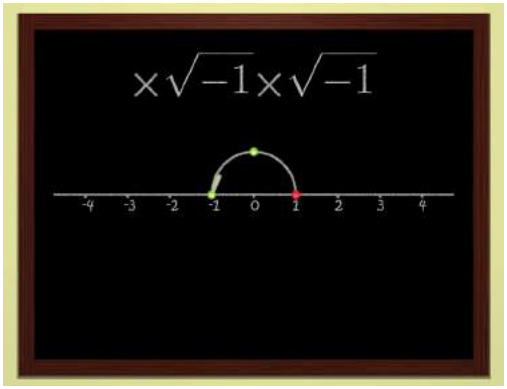


(a) $-2 \times -2 = 4$.

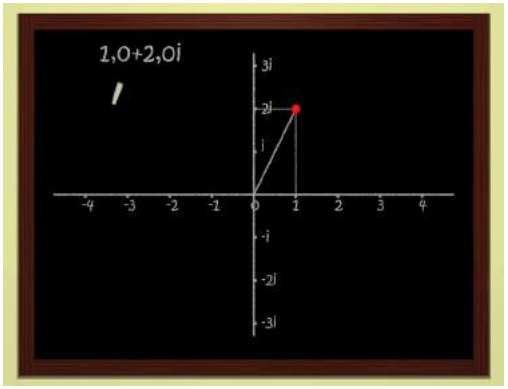


(b) La multiplication par -1 correspond à une rotation de 180° sur la droite réelle.

FIGURE 2.2 – Interpréter les nombres comme des *similitudes* de la droite réelle (dilatations + rotations) permet de nous ouvrir l'esprit : la multiplication n'est plus que la composition des fonctions.



(a) Les racines carrées de -1 (la rotation de 180°) sont alors toutes trouvées : ce sont les rotations d'angles 90° et 270°, que l'on note respectivement $+i$ et $-i$.



(b) Tout point du plan peut être associé à une unique similitude.

FIGURE 2.3 – Penser un point (x, y) comme "l'unique similitude qui envoie $(1, 0)$ sur (x, y)" permet de munir le plan d'une structure multiplicative. Miracle : celle-ci est compatible avec l'addition vectorielle, introduite au collège.

L'histoire de ces nouveaux nombres est bien longue et il n'est pas dans notre intention de décrire les étapes qui ont conduit à des bases solides. Il suffira de dire, pour simplifier à l'extrême, qu'au tournant du dix-neuvième siècle, quelques mathématiciens, dont Gauss, Wessel et Argand, ont pris conscience du caractère géométrique de ces nombres imaginaires. Le film montre une présentation simplifiée d'une idée très simple d'Argand.

Le nombre -1 est associé à la symétrie par rapport à l'origine sur la droite, c'est-à-dire à une rotation d'un demi-tour. Chercher une racine carrée pour -1, c'est chercher une transformation qui, effectuée deux fois de suite, serait une rotation d'un demi-tour. Argand déclare donc que la racine carrée de -1 doit être associée à la rotation d'un quart de tour, tout simplement. Faire deux rotations d'un quart de tour, c'est faire une rotation d'un demi-tour, c'est-à-dire multiplier par -1.

Si on part de cette idée, on a envie de dire que la racine carrée de -1 est obtenue à partir de 1 en tournant d'un quart de tour. Bien sûr, l'image de 1 par une rotation d'un quart de tour n'est pas sur la droite et nous venons de décider que la racine carrée de -1 est un point qui n'est pas sur la droite mais dans le plan !

L'idée est simple et jolie : considérer les points du plan comme des nombres. Alors bien sûr, ce ne sont plus les mêmes nombres que ceux auxquels nous sommes habitués. Pour cette raison, on dit que les nombres "traditionnels" sont les nombres réels, et les nombres que nous sommes en train de définir, associés aux points du plan, sont les nombres complexes.

Si nous repérons un point du plan par ses deux coordonnées (x, y), qui sont des nombres réels, la droite dont nous sommes partis est la droite d'équation $y = 0$, et le point qui est l'image de $(1, 0)$ par la rotation d'un quart de tour est $(0, 1)$. C'est donc ce point qu'Argand considère comme la racine carrée de -1. Les mathématiciens, toujours étonnés par ce "tour de passe passe", appellent ce point i, comme "imaginaire". Puisque nous voulons des nombres qu'on peut ajouter entre eux, on peut considérer le nombre $x + iy$: il lui correspond le point du plan de coordonnées (x, y).

En résumé, Argand nous incite à considérer les points (x, y) du plan non pas comme deux nombres (réels) mais plutôt comme un seul nombre (complexe). Cela peut sembler très étonnant et peut-être artificiel mais nous verrons que cette idée est très puissante.

Arithmétique complexe

La suite n'est pas difficile. Après toutes ces spéculations, on définit un nombre complexe z comme étant la donnée de deux nombres réels $(\mathbf{x}, \mathbf{y})$, c'est-à-dire un point du plan, et on le note $\mathbf{z = x + iy}$. Il s'agit ensuite de montrer qu'on peut ajouter ces nombres complexes, les multiplier, et aussi que toutes les propriétés du calcul auxquelles nous sommes habitués sont encore valides. Par exemple, il faut s'assurer que la somme des nombres complexes est la même quelle que soit l'ordre dans lequel on les ajoute. Tout cela peut être fait rigoureusement mais ce n'est bien sûr pas le but du film...

Pour l'addition c'est facile : on a la formule

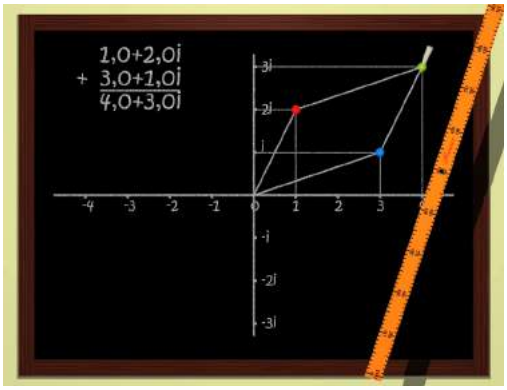
$$(x + iy) + (x' + iy') = (x + x') + i(y + y'), \tag{2.1}$$

si bien qu'ajouter des nombres complexes revient à ajouter des vecteurs. Pour la multiplication, c'est un peu plus difficile :

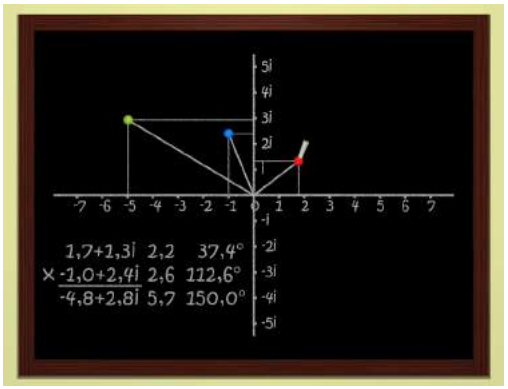
$$(x + iy) \cdot (x' + iy') = xx' + ixy' + iyx' + i2yy' \tag{2.2}$$

$$= (xx' - yy') + i(xy' + x'y), \tag{2.3}$$

mais ici, c'est par un petit miracle que cette formule est satisfaisante. Par exemple, il n'est pas du tout évident avec cette formule qu'on peut multiplier trois nombres complexes dans n'importe



(a) Additionner deux complexes, c'est procéder à une somme coordonnées à coordonnées. $(x_1 + iy_1) + (x_2 + iy_2) = (x_1 + x_2) + i(y_1 + y_2)$.



(b) Multiplier deux complexes, c'est faire le produit des modules et la somme des arguments. $r_1 e^{i\theta_1} \cdot r_2 e^{i\theta_2} = r_1 r_2 e^{i(\theta_1+\theta_2)}$.

FIGURE 2.4 – Le "plan" complexe $\mathbb{C}$ est muni d'une structure *additive*, donnée par l'addition des vecteurs, et d'une structure *multiplicative*, donnée par la composition des similitudes. Fait remarquable : cette structure géométrique donne l'unique loi $\times$ qui soit *compatible* avec $+$ (i.e. associative, distributive, commutative), étendant celle de $\mathbb{R}$ avec un élément symbolique i de carré -1. Elle avait été découverte, comme un simple jeu formel, par les mathématiciens du XVI^{e} siècle.

quel ordre pour trouver le même résultat, ou encore qu'on peut toujours diviser par un nombre non nul. Ce petit miracle n'est pas expliqué dans le film... cela nous aurait mené trop loin !

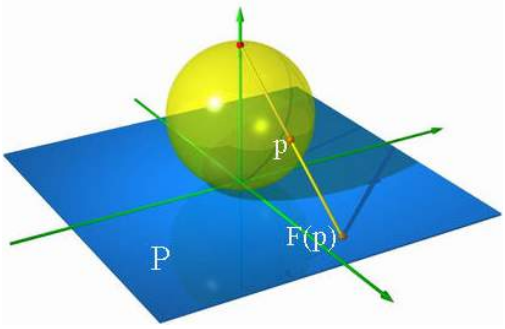
Deux notions seront utiles pour la suite :

Le *module* d'un nombre complexe $z = x+iy$ est simplement la distance du point correspondant (x, y) à l'origine. On le note $|z|$ et il est égal, d'après le théorème de Pythagore à $\sqrt{x^2 + y^2}$. Par exemple, le module de i est égal à 1 et celui de $1 + i$ à $\sqrt{2}$.

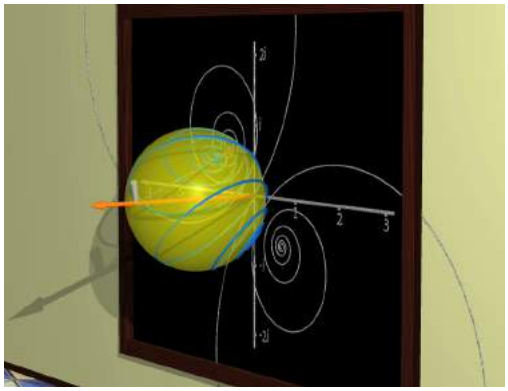
L'*argument* indique la direction de z. On le note $Arg(z)$ et ce n'est rien d'autre que l'angle entre l'axe des abscisses et la droite joignant l'origine à (x, y). Cet argument n'est défini que si z est non nul. Par exemple, l'argument de i est de 90 degrés, celui de 1 est nul, celui de -1 de 180 degrés, et celui de $1 + i$ de 45 degrés.

Les mathématiciens ont longtemps essayé de faire la même chose dans l'espace de dimension 3 : comment multiplier des points dans l'espace ? Il leur a fallu attendre longtemps avant de comprendre que ce n'est pas possible. Dans l'espace de dimension 4, ils ont découvert que c'était partiellement possible, à condition d'abandonner l'idée que le multiplication vérifie $ab = ba$! et ils ont fini par découvrir qu'en dimension 8, c'est encore possible, à condition d'abandonner l'idée que $(ab)c = a(bc)$, avant de comprendre, au milieu du vingtième siècle, que dans les dimensions autres que 1,2,4 et 8, il n'y a vraiment aucun moyen de multiplier les points ! Pour comprendre quelque chose aux phrases mystérieuses qui précèdent, on pourra lire les articles Wikipédia sur les nombres hypercomplexes, les quaternions et les octonions.

En résumé, les points du plan sont définis par un seul nombre... complexe. Le plan que nous avions dit être de dimension 2 est maintenant de dimension 1 ! Il n'y a bien sûr pas de contradiction : le plan est de dimension 2 réelle mais c'est une droite de dimension 1 complexe. Plan réel, droite complexe... Dimension 2 réelle, dimension 1 complexe. Jeu de mots ?

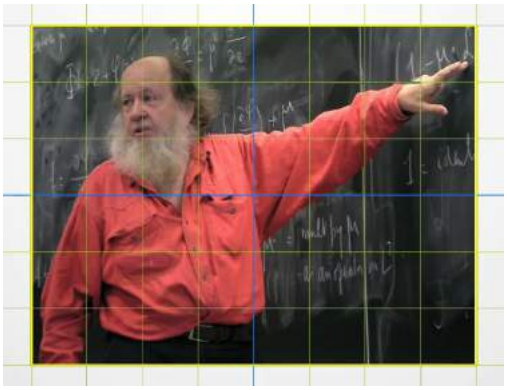


(a) La projection stéréographique définit une bijection entre la sphère et le plan, avec un "point à l'infini" associé au pôle Nord.

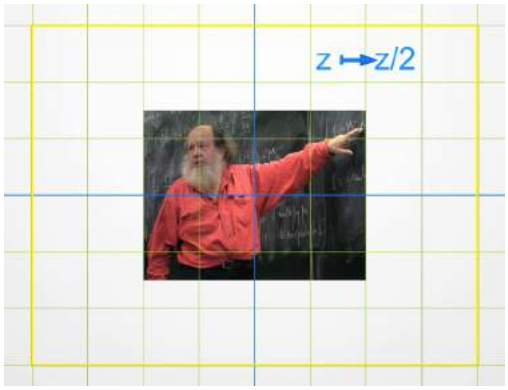


(b) La sphère est une droite projective complexe, compacte.

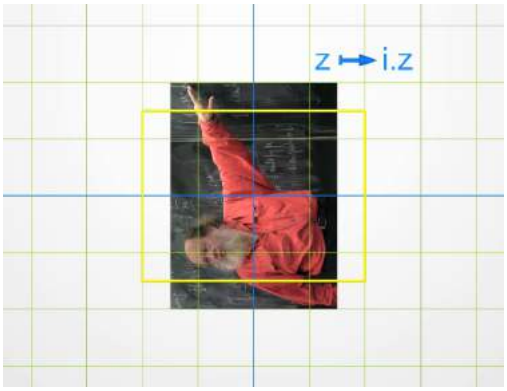
FIGURE 2.5 – La projection stéréographique permet de comprendre la sphère $\mathbb{S}^2$ comme une droite complexe, munie d'un point à l'infini. De nombreuses notions algébriques définies a priori sur $\mathbb{C}$ prendront en fait tout leur sens sur la sphère, qui traite l'infini comme un point "à part entière" : on pense notamment aux polynômes, aux coniques (ellipses-paraboles-hyperboles).



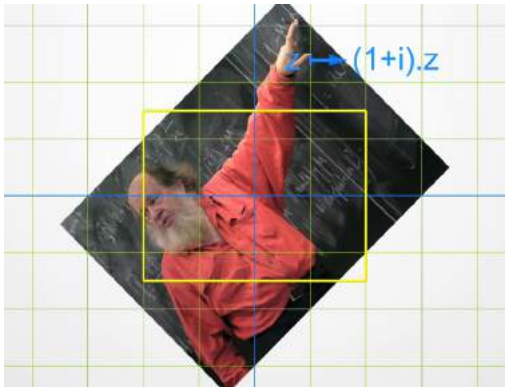
(a) Transformation identité, $z \mapsto z$.



(b) Homothétie, $z \mapsto z/2$.



(c) Rotation, $z \mapsto iz$.



(d) Similitude, $z \mapsto (1+i)z$.

FIGURE 2.6 – Premières transformations du plan complexe : les similitudes.

... encore la projection stéréographique !

Rappelez-vous la projection stéréographique ; elle transforme la sphère de dimension 2, privée du pôle nord, sur le plan tangent au pôle sud – dans le cours, nous la découvrirons section 4.4.1. Si un point s'approche du pôle nord, sa projection s'éloigne dans le plan si bien qu'on dit qu'elle tend vers l'infini. On dit d'ailleurs parfois que le pôle nord est le point à l'infini.

Maintenant, si on pense au plan tangent au pôle sud comme une droite complexe, on comprend pourquoi la sphère de dimension 2 (réelle !) est souvent qualifiée de droite projective complexe. Voilà un bel exemple d'acrobatie mathématique : appeler droite une sphère !

Henri Poincaré ne disait-il pas que les mathématiques consistent à donner le même nom à des choses différentes ?

Transformations

Le chapitre 6 du film propose de donner un peu d'intuition sur les nombres complexes à travers certaines transformations de la droite complexe.

Une transformation T est une opération qui associe à chaque nombre complexe z, c'est-à-dire à chaque point du plan, un autre point $T(z)$. Pour l'illustrer, on place le portrait d'Adrien Douady dans le plan et on montre ensuite son image par la transformation : chaque pixel qui constitue le portrait est transformé par T.

Adrien choisit plusieurs exemples de transformation T :

$\mathbf{T(z) = z/2}$ Chaque nombre est divisé par deux. Bien sûr, l'image est réduite deux fois : un zoom arrière ! On appelle cela une *homothétie*.

$\mathbf{T(z) = iz}$ Il s'agit simplement d'une rotation d'un quart de tour, par définition de i...

$\mathbf{T(z) = (1 + i)z}$ Puisque le module de $1 + i$ est $\sqrt{2}$ et son argument est 45 degrés, il s'agit de composer une rotation de 45 degrés et une homothétie d'un facteur $\sqrt{2}$. On appelle cela une similitude. C'est l'un des grands avantages des nombres complexes : ils permettent d'écrire très simplement les similitudes comme des multiplications.

$\mathbf{T(z) = z^2}$ Voilà notre première transformation non linéaire. En plaçant la photo en deux endroits différents, on peut prendre conscience de l'effet du passage au carré dans la droite complexe : les modules sont élevés au carré et les arguments sont doublés.

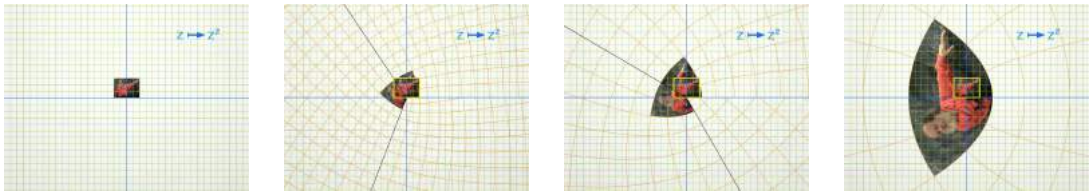
$\mathbf{T(z) = -1/z}$ Il s'agit d'une transformation proche de celle qu'on appelle d'ordinaire l'inversion. Bien sûr, l'origine qui correspond au nombre 0, ne peut pas être transformée mais on convient de dire qu'elle est envoyée à l'infini. La raison est très simple : si un nombre complexe z s'approche de 0, c'est-à-dire si son module tend vers 0, son transformé $-1/z$ a un module qui est l'inverse de celui de z et qui tend donc vers l'infini. La transformation a donc la propriété d'"exploser", c'est-à-dire de transporter très loin les petits voisinages de l'origine, jusqu'à sortir hors de l'écran... Réciproquement, les points qui sont très éloignés de l'origine sont "écrasés" très près de l'origine.

Pendant très longtemps, les manuels scolaires donnaient une très grande importance à l'inversion qui permet de démontrer de bien jolis théorèmes. La propriété principale de l'inversion est qu'elle transforme les cercles en des cercles ou des droites. Les artistes utilisent souvent ce genre de transformations et leur donnent le nom d'anamorphose.

Plus généralement, si on choisit quatre nombres complexes a,b,c,d, on peut considérer la transformation

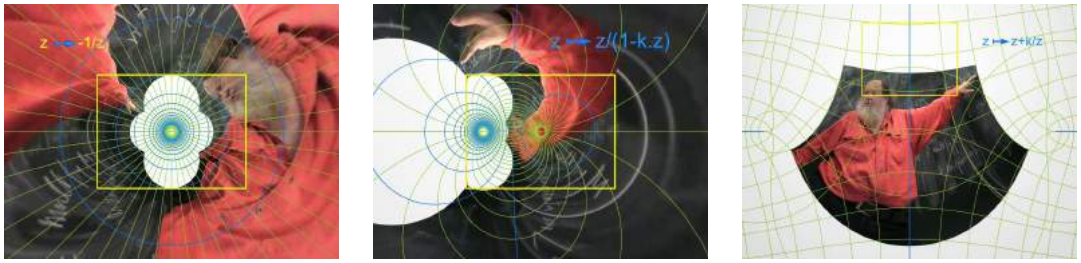
$$T(z) = \frac{az + b}{cz + d}. \tag{2.4}$$

Ces transformations portent plusieurs noms en mathématiques : transformations de Moebius, homographies, transformations projectives, mais leur propriété principale est d'envoyer les cercles



(a) Transformation identité, $z \mapsto z$. (b) Entre-deux. (c) Entre-deux. (d) Monôme, $z \mapsto z^2$.

FIGURE 2.7 – Passage de l'identité au monôme $z \mapsto z^2$. À l'exception de 0, chaque point a exactement deux antécédents : le passage au carré est un *revêtement à deux feuillets* de $\mathbb{C}^\star$ par lui-même. On retrouvera cette transformation à la Figure 2.17.



(a) $z \mapsto -1/z$. (b) $z \mapsto z/(1 - k \cdot z)$. (c) $z \mapsto z + k/z$.

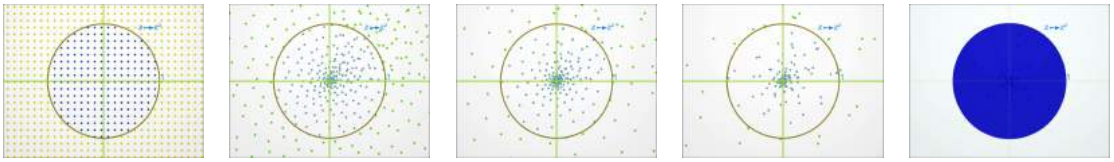
FIGURE 2.8 – Quelques homographies proposées dans le film.

sur des cercles ou des droites. Ce groupe de transformations est celui d'une magnifique géométrie appelée circulaire, proche de la géométrie non euclidienne, mais c'est une autre histoire – qui sera détaillée au chapitre 4 de notre polycopié !

Dans le film, on s'intéresse à la transformation

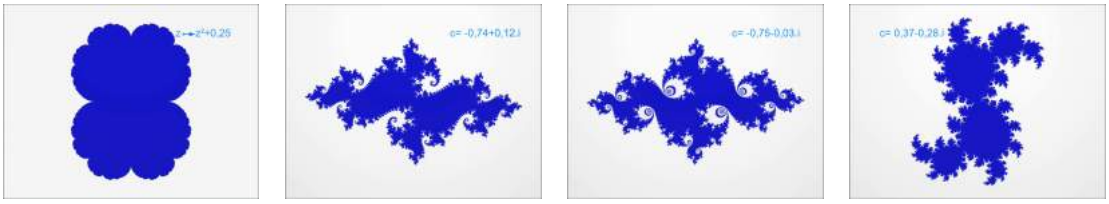
$$T(z) = z + \frac{k}{z}. \tag{2.5}$$

Cette transformation a été étudiée par Joukovski dans ses études sur l'aérodynamique des ailes d'avions ! Mais Adrien Douady aurait pu choisir d'autres transformations, en particulier qui lui donnent une ligne plus fine que celle-ci ! Le but de cette illustration est de montrer une propriété fondamentale de ce type de transformations. Bien sûr, elles ne transforment plus les cercles en cercles, seules les transformations de Moebius le font ; mais cela est vrai au niveau infinitésimal. Si on prend un petit cercle et on considère la courbe transformée, elle n'est pas un cercle mais elle est très proche d'un cercle, d'autant plus proche que le cercle initial est petit. Une autre manière d'exprimer la même chose est de dire que les transformations en question se comportent comme des similitudes au niveau infinitésimal. Ces transformations sont appelées *holomorphes* ou *conformes*. Les racines grecque et latine "holo" et "con" signifient "même", et morphe signifie bien sûr "forme" : autrement dit ces transformations préservent les formes. L'étude des fonctions holomorphes est l'un des chapitres les plus importants des mathématiques, comme nous aurons l'occasion de le voir à la fin du chapitre.



(a) État initial. (b) Première... (c) Deuxième... (d) Troisième... (e) Julia.

FIGURE 2.9 – L'ensemble de Julia associé au monôme $z \mapsto z^2+0$ est le disque unité. On applique le polynôme une fois, deux fois, trois fois... À la limite, seuls les points de module $|z| \leqslant 1$ conservent une orbite bornée.



(a) $z \mapsto z^2 + .25$ (b) $z \mapsto z^2 - .74 + .12i$ (c) $z \mapsto z^2 - .75 - .03i$ (d) $z \mapsto z^2 + .37 - .28i$

FIGURE 2.10 – Lorsque la constante c est non nulle, la dynamique se complexifie, ce qui est reflété dans l'ensemble de Julia associé.

Dynamique holomorphe

Dans la deuxième partie du chapitre 6, Adrien Douady propose une initiation à un magnifique sujet d'étude auquel il a apporté des contributions majeures. Il s'agit de l'étude des ensembles de Julia, qui outre leur intérêt mathématique fondamental ont une beauté extraordinaire (et les deux choses sont bien sûr liées). Il est rare qu'une théorie mathématique puisse être illustrée d'une manière si belle et de nombreux artistes se sont inspirés de ces images.

L'idée de départ est très simple : on choisit un nombre complexe c quelconque. Puis, on considère la transformation $T_c(z) = z^2 + c$. Il s'agit donc dans un premier temps d'élever au carré un nombre puis de le translater en lui ajoutant c. Partant d'un point initial z, son transformé est un point $z_1 = T_c(z)$, puis on considère le transformé du transformé $z_2 = T_c(z_1)$ et on continue à l'infini en produisant une suite de nombres complexes z_n dont chacun est le transformé du précédent. On dit que la suite z_n est l'orbite du point initial z par la transformation T_c. Etudier le comportement de cette suite z_n, c'est comprendre la dynamique de T_c. Il s'agit bien sûr d'un exemple très simple, mais cet exemple est suffisamment riche pour engendrer de très belles mathématiques.

Considérons d'abord le cas où $c = 0$. Il s'agit alors d'effectuer de manière répétée la transformation $T_c(z) = z^2$. Le module de chaque z_n est donc le carré du précédent. Si le module de z est inférieur à 1, c'est-à-dire si z est à l'intérieur du disque de rayon 1 centré sur l'origine, tous les z_n vont rester dans ce disque. Par contre si le module de z est strictement supérieur à 1, les modules des z_n vont croître sans cesse et même tendre vers l'infini : l'orbite de z va finir par quitter l'écran !

Dans le premier cas, on dit que l'orbite est stable : elle reste dans une zone limitée du plan. Dans le second cas, elle est instable : elle fuit vers l'infini. L'ensemble des points z dont l'orbite est stable est donc le disque.

De manière générale, pour chaque valeur de c, on peut aussi distinguer deux sortes de points z.

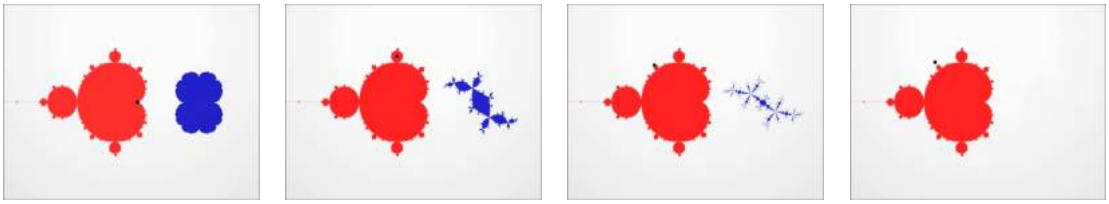


FIGURE 2.11 – Un point c (en noir) est dans l'ensemble de Mandelbrot (en rouge) si le Julia associé à $z \mapsto z^2 + c$ (en bleu) est connexe, visible à l'écran. C'est le cas pour les trois premières images, mais pas pour la quatrième.

L'orbite de z par T_c peut être stable, si elle reste dans une partie limitée du plan, ou instable dans le cas contraire. L'ensemble des z dont l'orbite est stable est appelé l'ensemble de Julia rempli de la transformation T_c. Comprendre la structure de ces ensembles de Julia et la manière dont ils varient lorsque c varie est un enjeu majeur de la théorie des systèmes dynamiques holomorphes. Dans un premier temps, Adrien Douady nous montre quelques exemples d'ensembles de Julia pour diverses valeurs de c. Certains portent des noms exotiques, comme par exemple le lapin (voyez-vous ses oreilles ?) pour $c = -0.12 + 0.77i$.

On sait depuis le début du vingtième siècle que l'ensemble de Julia rempli peut être de deux types. Il peut être, comme dans les exemples montrés précédemment, d'un seul tenant, connexe comme on dit en mathématiques, ou bien il peut être totalement discontinu, constitué d'une infinité de morceaux éclatés, chacun d'entre eux étant d'intérieur vide, ce qui veut dire en clair qu'on ne les voit pas sur un dessin ! Par conséquent, il y a des valeurs de c pour lesquelles on voit l'ensemble de Julia et d'autres pour lesquelles on ne le voit pas (même si bien sûr il est présent). L'ensemble des valeurs de c pour lesquelles on voit bien l'ensemble de Julia (pour lesquels l'ensemble de Julia est connexe) est appelé l'ensemble de Mandelbrot, pour rendre hommage à Benoît Mandelbrot, son inventeur. Adrien Douady a beaucoup travaillé pour comprendre cet ensemble ; il a par exemple contribué à montrer qu'il est lui-même connexe et il aurait bien aimé (comme beaucoup d'autres) montrer qu'il est localement connexe...

La fin du chapitre est consacrée à une plongée dans l'ensemble de Mandelbrot, plongée profonde puisque le facteur de dilatation est de l'ordre de deux cent milliards ! On peut observer cette scène de deux manières. On peut la regarder et l'admirer tout simplement : c'est suffisamment joli pour cela ! Mais on peut aussi se poser quelques questions...

Par exemple, quelle est la signification des couleurs ? Un ancien théorème affirme que l'ensemble de Julia de T_c n'est pas connexe, autrement dit que c n'est pas dans l'ensemble de Mandelbrot, si et seulement si l'orbite de 0 par T_c est instable. Pour une valeur de c donnée, on peut donc prendre l'orbite de $z = 0$ par T_c et observer son comportement pour les grandes valeurs de n. Si z_n devient très grand rapidement, c'est que c n'est pas dans l'ensemble de Mandelbrot et même qu'il en est assez éloigné. Si la suite z_n tend vers l'infini mais plus lentement, le point c n'est toujours pas dans l'ensemble de Mandelbrot mais il en est en quelque sorte plus proche. La couleur avec laquelle on colorie le point c dépend de la vitesse de fuite vers l'infini de l'orbite z_n, montrant ainsi la "proximité" à l'ensemble de Mandelbrot. Si par contre z_n reste dans une zone limitée, alors c est dans l'ensemble de Mandelbrot et on le colorie en noir.

L'ensemble de Mandelbrot sur la figure ci-dessus à été colorié de cette façon, mais il existe des dizaines de méthodes. Dans le film, on a utilisé la méthode dite "Inégalité du triangle" : lorsque le module de z_n devient plus grand qu'une certaine valeur, on calcule les modules $A = |z_n - z_{n-2}|$, $B = |z_n - z_{n-1}|$ et $C = |z_{n-1} - z_{n-2}|$. $A/(B + C)$ donne toujours un résultat entre 0 et 1, et on utilise ce résultat pour indiquer la position sur une palette de couleurs.

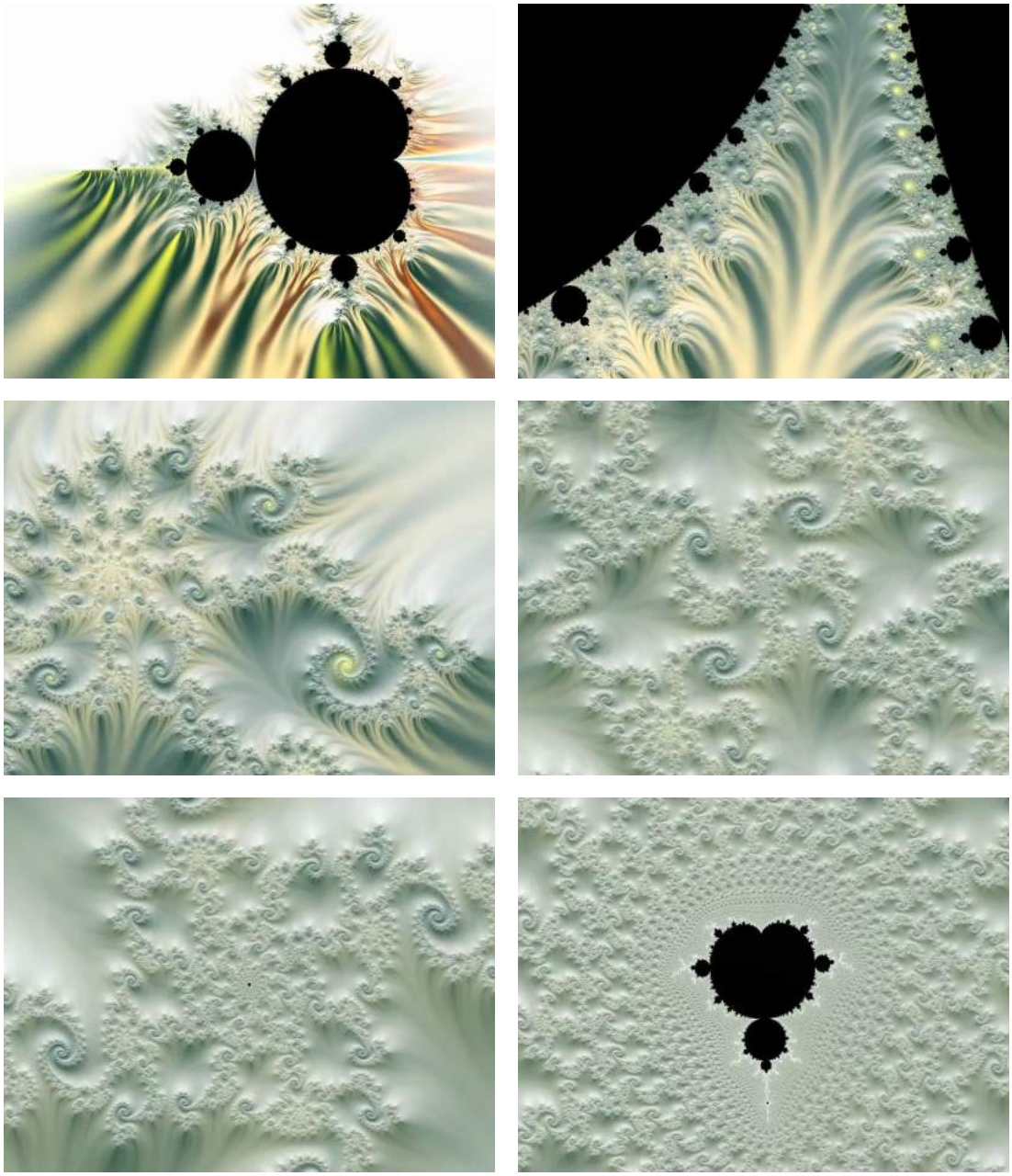


FIGURE 2.12 – Plongée dans l'ensemble de Mandelbrot.

Pourquoi à certains moments a-t-on l'impression de voir apparaître de nouvelles petites copies noires de l'ensemble de Mandelbrot ? C'est beaucoup plus difficile à expliquer et c'est l'une des découvertes importantes d'Adrien Douady : l'ensemble de Mandelbrot possède des propriétés d'autosimilarité : une caractéristique fréquente des ensembles fractals.

Polynômes et théorème fondamental de l'algèbre

Cette introduction aux nombres complexes nous aura permis de démystifier l'existence et les propriétés de $\mathbb{C}$ de manière fort convaincante. Surtout, grâce aux réalisateurs du film *Dimensions*, nous avons pu profiter de notre petit voyage pour entrevoir ce beau sujet qu'est l'étude des systèmes dynamiques (holomorphes).

Pour terminer ce chapitre, je voudrais maintenant revenir à des bases solides, des résultats accessibles à des élèves non-mathématiciens. Vous montrer que les belles idées se cachent même derrières les théorèmes fondamentaux du lycée.

Polynômes complexes On s'en souvient : les "polynômes de degré 1 et 2" occupent une place importante dans les programmes du secondaire... Mais qu'est-ce, au juste, qu'un polynôme P ? Très simplement : une somme pondérée d'applications "puissances", i.e. la donnée d'une suite finie de coefficient p_d, ..., p_0 dans $\mathbb{C}$, avec

$$P(z) = \sum_{k=0}^{d} p_k z^k = p_d z^d + \cdots + p_2 z^d + p_1 z^1 + p_0 z^0. \tag{2.6}$$

On appellera *degré* du polynôme P le plus grand indice associé à un coefficient non nul. Un polynôme de degré 0 sera donc simplement identifié à une constante p_0, tandis qu'un polynôme de degré 1 correspondra à une application linéaire, et un polynôme de degré 2, à une application quadratique :

$$P(z) = p_2 z^2 + p_1 z + p_0 \tag{2.7}$$

$$= az^2 + bz + c \quad \text{avec les notations du lycée.} \tag{2.8}$$

Les polynômes sont donc, en un sens, les expressions algébriques les plus simples ; les généralisations immédiates des problèmes linéaires que nous avions étudié au chapitre précédent. Le théorème que je vous propose d'étudier est alors le suivant :

Théorème 2.1 (D'Alembert, Gauss)**.** *Soit P un polynôme complexe non constant – i.e. de degré supérieur ou égal à 1. Alors l'équation*

$$P(z) = 0 \tag{2.9}$$

admet au moins une solution dans $\mathbb{C}$.

Intérêt Ce "théorème fondamental de l'algèbre" mérite bien son nom. En garantissant l'existence d'une solution complexe pour toute équation polynomiale, il justifie la construction de $\mathbb{C}$: envisagé au départ comme une simple extension de $\mathbb{R}$, $\mathbb{C}$ devient en fait *le* cadre privilégié pour faire des calculs algébriques.

Notez que ce résultat est a priori surprenant : en ajoutant les seules racines $\pm i$ du polynôme $X^2 + 1$ – et leurs multiples/sommes avec des réels – nous avons résolu le problème des solutions pour *toute* équation polynomiale.

La propriété qu'a $\mathbb{C}$ d'être *algébriquement clos* est en fait comparable à la *complétude* de $\mathbb{R}$ dont nous avions discuté au chapitre précédent. Rappelez-vous : en garantissant la propriété des *valeurs intermédiaires*, la complétude de la droite réelle nous avait permis de donner une solution à tout problème d'intersection de courbes, de trajectoires. Si l'instant de rencontre entre une balle et le sol était impossible à définir comme un rationnel, il était par contre tout à fait légitime en tant que *nombre réel* : c'est donc que $\mathbb{R}$ est le cadre naturel pour faire de la mécanique classique, le cadre où on peut appeler un instant par son *nom* sans périphrases encombrantes.

Pour les nombres complexes, c'est la même chose. Le théorème 2.1 assure simplement que si l'on étudie des équations polynomiales – elles apparaissent naturellement en physique, en géométrie, en électronique ou partout ailleurs – alors il n'y a pas de mauvaise surprise à craindre. Les solutions de nos équations seront peut-être *complexes*, et éventuellement impossibles à atteindre physiquement ; mais au moins, nous pourrons en parler, les étudier, les cataloguer. L'instant du choc $t_{\text{choc}} = \sqrt{2}\,\text{s}$ était irrationnel à la page 200 ? C'est donc qu'il est impossible de le "prendre en photo" à la bonne fraction de seconde ; mais cela ne devrait pas nous empêcher d'en parler.

J'espère vous avoir convaincu de la pertinence de ce résultat. Plutôt que de m'appesantir sur les propriétés des nombres complexes et leurs applications à la vie courante – ce que nous ferons au chapitre 3 – je voudrais consacrer la deuxième moitié de ce chapitre à une discussion autour de la notion de *théorème*.

Variété des approches Plus que le résultat en lui-même, connu et démontré depuis la fin du XVIII$^{\text{e}}$ siècle, ce sont ses *preuves* qui intéressent le mathématicien moderne. Mais à quoi bon vérifier, re-vérifier et re-revérifier un énoncé si célèbre ? C'est que contrairement à ce que l'on croit souvent, le rôle fondamental du mathématicien n'est pas de *vérifier*, mais de *comprendre*.

Bien connaître le travail du mathématicien, c'est réaliser que plus qu'une liste de *théorèmes* à utiliser comme boîtes noires, un (bon) article de recherche est la proposition d'un nouveau *point de vue* sur une question, ouverte ou non. Bien sûr, les "grands problèmes", les "grandes conjectures" ont toujours attiré les mathématiciens ingénieux. Mais loin des images de trésors enfouis au petit bonheur sous le sable d'une plage, il faut se les représenter comme perchés à des hauteurs si élevées qu'elles semblent inaccessibles : seul arrivera à décrocher la Lune celui qui construira la première fusée, ouvrant ainsi la voie à une nouvelle ère de découvertes.

Prenez par exemple le problème de Syracuse. Partant d'un entier n quelconque, il s'agit de lui appliquer itérativement la règle suivante :

« S'il est pair, divisez-le par deux. Sinon, multipliez-le par trois, et ajoutez 1. » (2.10)

On peut ainsi construire les "parties" :

$$n = 1 \to\ 4 \to\ 2 \to\ 1 \to \cdots \tag{2.11}$$

$$n = 3 \to 10 \to\ 5 \to 16 \to\ 8 \to\ 4 \to\ 2 \to\ 1 \to \cdots \tag{2.12}$$

$$n = 7 \to 22 \to 11 \to 34 \to\ 17 \to 52 \to 26 \to 13 \to\ 40 \to 20 \to 10 \to 5 \to 16 \to 8 \to 4 \to 2 \to 1 \to \cdots \tag{2.13}$$

En fait, pour tout choix de l'entier initial n, on finit par boucler sur le cycle élémentaire $1 \to\ 4 \to\ 2 \to\ 1 \to\ \cdots$ – ç'a été testé jusqu'à cinq milliards de milliards... Mais aussi incroyable que celui puisse paraître, depuis que cette petite propriété a été remarquée en 1928, *personne* n'a jamais réussi à la démontrer. Et attention, ce n'est pas faute d'efforts : dans les années 60, tant d'énergie avait été dépensée à ce sujet qu'on en arrivait même à dire en plaisantant que le problème n'était autre... qu'une invention soviétique destinée à détourner l'effort de recherche capitaliste !

Non, si aucune preuve de la "conjecture $3x + 1$" n'a pu être trouvée, c'est bien, d'après le grand Paul Erdös, parce que « les mathématiques ne sont pas encore prêtes pour ce genre de problèmes ». Alors, quelle gloire pour celui qui saura faire avancer la science jusqu'à ce point !

Preuves du théorème fondamental de l'algèbre Pour la question qui nous intéresse ici, les honneurs sont traditionnellement partagés entre Jean le Rond D'Alembert – celui de l'encyclopédie – qui en proposa une preuve incomplète en 1746, et Carl Friedrich Gauss qui en trouva pas moins de quatre au cours de sa vie, de sa thèse de doctorat en 1799 à un travail tardif en 1849 – six ans avant sa mort.

Exposition du plan Dans les pages qui suivent, nous allons voir que chaque preuve originale du théorème repose sur une nouvelle propriété, une nouvelle manière de comprendre la notion de "polynôme". Sans plus attendre, je vous propose donc d'attaquer notre programme par une preuve élémentaire, *directe*. Peu à peu, au fil des pages, une intuition claire de ce qu'*est* un polynôme complexe devrait se former dans votre esprit : espérons donc que le voyage vous plaira !

Preuve directe

$|P|$ atteint son minimum Soit P un polynôme complexe de degré supérieur ou égal à 1. On commence par remarquer deux choses :

1. P est continu de $\mathbb{C}$ dans $\mathbb{C}$, i.e.

$$P(z_0 + h) \xrightarrow{h \to 0} P(z_0), \tag{2.14}$$

   car c'est une somme finie de fonctions continues.

2. P diverge à l'infini, i.e.

$$|P(z)| \xrightarrow{|z| \to +\infty} +\infty, \tag{2.15}$$

   car, avec d le degré de P, on a

$$|P(z)| = |z|^d \cdot \underbrace{\left| p_d + \frac{p_{d-1}}{z} + \cdots + \frac{p_0}{z^d} \right|}_{\longrightarrow |p_d|}. \tag{2.16}$$

On déduit de cela que l'application *module* de P,

$$|P\,| : z \mapsto |P|(z), \tag{2.17}$$

atteint bien son *minimum* m en un point a de $\mathbb{C}$. Autrement dit, on a

$$|P(a)| = m \qquad \text{et} \qquad \forall\, z \in \mathbb{C},\ \ |P(z)| \geqslant m \geqslant 0. \tag{2.18}$$

En toute rigueur, démontrer l'existence d'un tel antécédent a au minimum m de $|P\,|$ n'est pas une chose "facile" : comme le théorème des valeurs intermédiaires, c'est un point qui requiert d'utiliser à fond la *complétude* de $\mathbb{R}$, puis de $\mathbb{C}$. Une démonstration de ces deux propriétés des fonctions à valeurs réelles sera proposée au chapitre B, qui introduit le vocabulaire *topologique* adéquat. En attendant, nous nous contenterons de l'admettre comme en classe de seconde : une fonction *continue* qui diverge vers $+\infty$ aux bords de son domaine de définition atteint nécessairement son minimum.

Preuve par l'absurde Supposons donc maintenant que P n'a pas de racine, qu'il n'existe pas de racine complexe z de P telle que $P(z) = 0$. En particulier, on a

$$m \ = \ |P(a)| \ > \ 0 \tag{2.19}$$

Normalisation du problème Plutôt que de s'encombrer de valeurs quelconques pour a et m, on simplifie les calculs ultérieurs en les *renormalisant* : quitte à considérer le polynôme

$$Q(z) = \frac{1}{P(a)} P(z - a), \tag{2.20}$$

on peut supposer que $a = 0$, $m = 1 = P(a)$. On peut donc écrire, avec $P_d \neq 0$,

$$P(z) = p_d z^d + \cdots + p_2 z^2 + p_1 z + 1, \tag{2.21}$$

et il s'agit de démontrer que l'hypothèse suivante est absurde :

$$\forall\ z \in \mathbb{C},\quad |P(z)|\ \geqslant\ 1. \tag{2.22}$$

P est localement surjectif On sait par hypothèse sur P (degré d non nul) qu'au moins un des coefficients p_k de l'écriture (2.21) est non nul : on va raisonner sur l'indice du plus petit d'entre eux, noté n, entier compris entre 1 et d, et on écrit simplement

$$P(z) = 1 + z^n \cdot \big(p_n + p_{n+1} z + \cdots + p_d z^{d-n} \big)\,. \tag{2.23}$$

En notant θ un argument de p_n dans $[-\pi, \pi]$ tel que $p_n = \rho e^{i\theta}$ (avec $\rho > 0$), et

$$z_t = t\, e^{i(\pi-\theta)/n}, \tag{2.24}$$

on constate alors que

$$P(z_t) \;=\; 1 + \qquad t^n e^{i(\pi-\theta)} \cdot \left(p_n + t\, p_{n+1}\, e^{i(\pi-\theta)/n} + \cdots + t^{d-n}\, p_d\, e^{i(\pi-\theta)(d-n)/n} \right) \tag{2.25}$$

$$\;=\; 1 + t^n e^{i(\pi-\theta)} \cdot \rho e^{i\theta} \cdot \left(1 \;+ t\, \frac{p_{n+1}}{p_n}\, e^{i(\pi-\theta)/n} + \cdots + t^{d-n}\, \frac{p_d}{p_n}\, e^{i(\pi-\theta)(d-n)/n} \right) \tag{2.26}$$

$$\;=\; 1 - \qquad \rho\, t^n \cdot \left(1 \;+ \underbrace{t\, \frac{p_{n+1}}{p_n}\, e^{i(\pi-\theta)/n} + \cdots + t^{d-n}\, \frac{p_d}{p_n}\, e^{i(\pi-\theta)(d-n)/n}}_{\xrightarrow{t\to 0}\ 0} \right). \tag{2.27}$$

Autrement dit, pour t suffisamment petit, $P(z_t)$ est légèrement à gauche de 1, proche de l'axe des réels, et donc de module strictement inférieur à 1. Ceci rentre en contradiction avec l'hypothèse de l'équation (2.22) : le théorème est donc démontré.

Une figure pour l'année prochaine... ! ! !

FIGURE 2.13 – Un polynôme P non constant explose à l'infini, et reste continu : son module atteint donc son minimum. Supposer par l'absurde que ce minimum est non nul, c'est supposer qu'il existe un disque D centré en 0 rayon m tel que $P(z)$ reste toujours en dehors de D, avec toutefois un certain $P(a)$ sur le bord du disque. Quitte à renormaliser, on peut supposer que $a = 0$, $m = 1$, $P(a) = P(0) = 1$. Mais alors, il est facile de trouver une direction, un angle $\theta' = (\pi - \theta)/n$ tel que $P(te^{i\theta'})$ approche $P(0) = 1$ par la gauche, quand t tend vers 0. Pour t suffisamment petit, on trouve donc un bon $P(z_t)$ dans le disque D, ce qui contredit notre hypothèse.

Preuve par homotopie

Après cette preuve *directe* conceptuellement simple sans être très élégante, retrouvons un concept que vous connaissez bien : celui d'*invariant*. Rappelez-vous : nous l'avions utilisé page 17 pour résoudre le jeu de pavage du damier et le problème de Loyd. Ici, il ne sera pas question de parité du nombre d'inversions, mais de nombre d'*enroulements*.

Premières définitions Avant d'arriver à notre théorème, définissons un *lacet* comme une application *continue* du cercle S^1 dans le plan complexe privé du point 0 – voir Figure 2.14.

On dira que deux lacets l et m sont *homotopes*, ou joignables par une déformation continue, s'il existe une fonction continue

$$F : [0,T] \times S^1 \to \mathbb{C}\backslash\{0\} \tag{2.28}$$
$$(t,s) \mapsto F(t,s) \tag{2.29}$$

telle que

$$\forall\ s \in S^1,\ F(0,s) = l(s) \text{ et } F(T,s) = m(s). \tag{2.30}$$

Autrement dit, un *chemin* F entre l et m est la donnée d'une collection continue de lacets $F_t = F(t,\cdot)$ telle que $F_0 = l$ et $F_T = m$: ceux-ci sont par exemple représentés Figure 2.15a.

Il n'est pas difficile de se convaincre que la relation d'homotopie est bien celle qui correspond à l'idée d'une déformation "sans déchirures" : deux lacets l et m sont homotopes si et seulement si on peut déformer l'un en l'autre continûment, le chemin F jouant alors le rôle de "film cinématique" avec une image F_t par instant t de $[0,T]$ donné.

Comme pour le jeu de taquin, la question essentielle est la suivante :

« Tous les lacets sont-ils joignables entre eux ? »

Cas du plan Dans le plan complexe "tout entier", la réponse est affirmative : tout lacet l est joignable au lacet constant égal à 0. Il suffit pour cela de considérer l'homotopie de rétraction

$$F : [0,1] \times S^1 \to \mathbb{C} \tag{2.31}$$
$$(t,s) \mapsto (1-t)\, l(s) \tag{2.32}$$

qui réduit uniformément l sur le point origine du repère. Elle se trouve représentée Figure 2.15.

Cas du plan épointé Dans le cadre qui nous intéresse ici, celui de $\mathbb{C}\backslash\{0\}$, la réponse est plus difficile à trouver. Il semble en effet impossible de passer du lacet identité à un lacet constant : le point 0 étant infranchissable, notre expérience du monde nous dit bien qu'à moins de *déchirer* notre lacet, il restera toujours solidement *enroulé* autour de l'origine du repère. Si l'on admet disposer d'une définition générale du *nombre d'enroulement*, ou nombre de tours que fait un lacet autour du point 0, le résultat suivant nous tire d'affaire :

Théorème 2.2. *Le nombre d'enroulements est un invariant d'homotopie sur les lacets de* $\mathbb{C}\backslash\{0\}$.

Autrement dit : il est impossible d'enrouler un bracelet élastique autour d'une barre de métro. Attention : tout intuitif qu'il soit, le théorème 2.2 est un résultat *difficile*. Définir proprement le nombre d'enroulement d'un lacet quelconque est déjà beaucoup trop technique pour un simple cours de vulgarisation... Tandis que démontrer son invariance par homotopie demande des heures de travail à vos camarades qui suivent le cours de *Topologie Algébrique* !

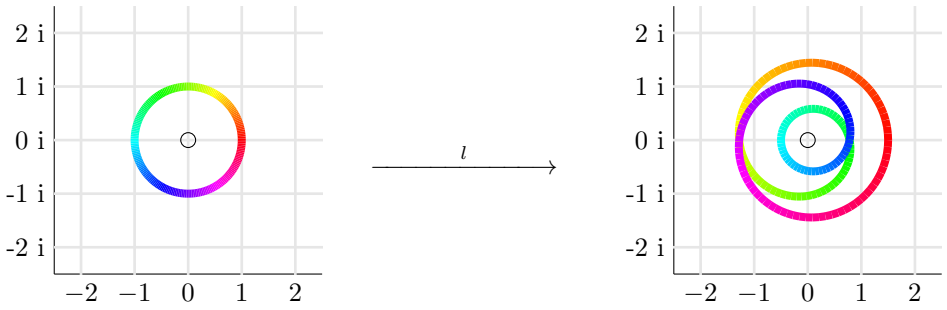
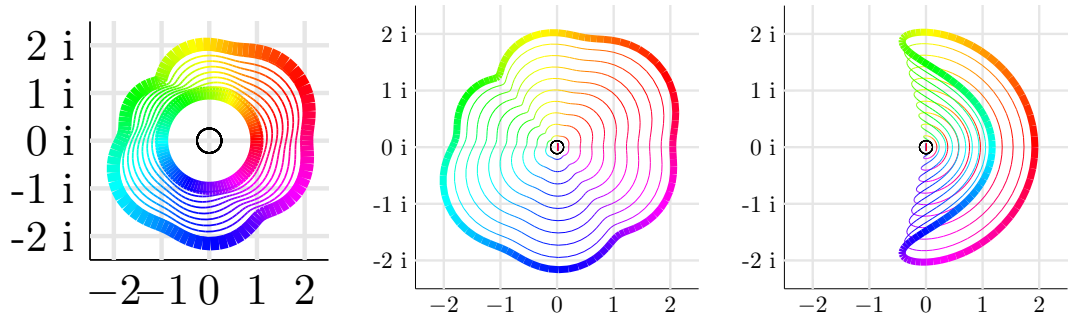


FIGURE 2.14 – Exemple de lacet : le cercle unité est envoyé continûment dans le plan complexe.

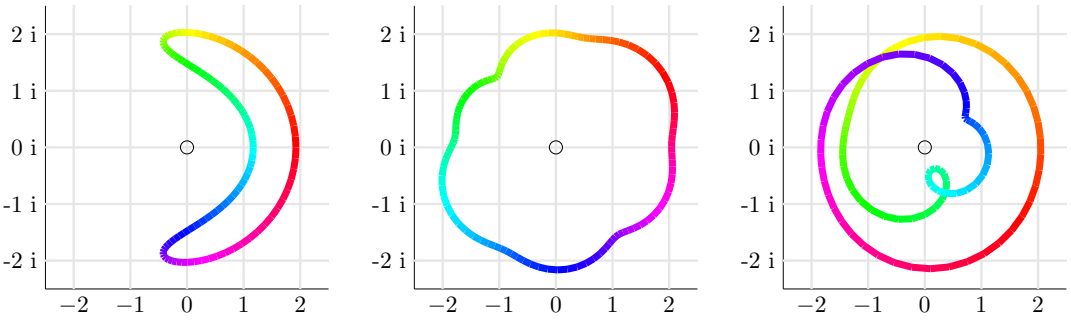


(a) Exemple d'homotopie entre le lacet identité et un lacet biscornu.

(b) Tout lacet du plan peut être rétracté sur le point 0...

(c) En composant les déformations, on peut donc joindre deux lacets quelconques sur $\mathbb{C}$.

FIGURE 2.15 – La relation d'homotopie illustrée dans le plan complet $\mathbb{C}$. On représente ici les étapes intermédiaires de la déformation pour $t = 0, 0.1, 0.2, \ldots, 1$.



(a) Nombre d'enroulements : 0.

(b) Nombre d'enroulements : 1.

(c) Nombre d'enroulements : 2.

FIGURE 2.16 – Le nombre d'enroulements étant invariant par homotopie dans le plan épointé, aucun de ces trois lacets ne peut-être relié à un des deux autres dans $\mathbb{C}\backslash\{0\}$.

Il est en effet impossible d'utiliser ici un schéma de preuve "exhaustif", en vérifiant pour chaque règle sa conservation de l'invariant : contrairement au jeu de taquin – 48 mouvements possibles – le "jeu de l'élastique" a une infinité de degrés de liberté ; une théorie neuve – celle des revêtements – doit donc être construite de toutes pièces. Pour ce cours, nous nous contenterons d'admettre cette propriété : il est impossible de joindre continûment deux lacets dont le nombre d'enroulements diffère.

Preuve du théorème fondamental de l'algèbre Revenons donc maintenant à notre polynôme P non constant, et supposons par l'absurde que P ne s'annule jamais. Le point clé de notre preuve sera de remarquer, comme à la Figure 2.17, qu'un polynôme de degré d fait "tourner" l'espace à l'infini en plus de le dilater.

Pour tout rayon r donné, on peut regarder l'image l_r par P du cercle rS^1 de centre 0 et de rayon r : par hypothèse, il s'agira d'un lacet à valeurs dans le plan *épointé* $\mathbb{C}\backslash\{0\}$. Illustré Figure 2.18, le lemme suivant est alors de première importance :

« Pour r assez grand, le nombre d'enroulements de l_r est égal au degré de P. »

Preuve du lemme Comme pour la preuve "directe", il suffit de prendre un rayon R tel que le terme de plus haut degré domine sur les autres : si on écrit

$$P(x) = p_d x^d + \cdots + p_1 x + p_0 \tag{2.33}$$

avec $p_d \neq 0$, il suffira de choisir R supérieur à 1 et à $(|p_{d-1}| + \cdots + |p_0|)/|p_d|$ pour obtenir pour tout angle θ la domination

$$|p_d \cdot (Re^{i\theta})^d| \geqslant 2 \cdot \left|p_{d-1} \cdot (Re^{i\theta})^{d-1} + \cdots + p_1 \cdot Re^{i\theta} + p_0\right|, \tag{2.34}$$

$$\text{i.e.} \quad |p_d \cdot (Re^{i\theta})^d| \geqslant 2 \cdot \left|P(Re^{i\theta}) - p_d \cdot (Re^{i\theta})^d\right|. \tag{2.35}$$

Le lacet $l_R : e^{i\theta} \in S^1 \mapsto P(Re^{i\theta}) \in \mathbb{C}\backslash\{0\}$ est donc quasi-équivalent au lacet

$$m_d : e^{i\theta} \in S^1 \mapsto p_d R^d e^{id\theta} \in \mathbb{C}\backslash\{0\}, \tag{2.36}$$

qui fait exactement d tours autour de 0. Plus formellement, on écrira simplement que

$$G : [0,1] \times S^1 \to \mathbb{C}\backslash\{0\} \tag{2.37}$$

$$(t,s) \mapsto t \cdot m_d(s) + (1-t)\, l_R(s) \tag{2.38}$$

est une homotopie bien définie de l_R vers m_d, qui ne s'annule pas grâce à l'équation (2.35) : en passant par le théorème 2.2, on obtient donc une preuve rigoureuse du fait que le nombre d'enroulements de l_R est égal au degré de P ; Cqfd.

Pour clore la preuve du théorème, il suffit alors de constater que l'application

$$l : [0,R] \times S^1 \to \mathbb{C}\backslash\{0\} \tag{2.39}$$

$$(r, e^{i\theta}) \mapsto P(re^{i\theta}) \tag{2.40}$$

est une homotopie du lacet constant égal à $P(0)$ vers le lacet l_R, comme illustré Figure 2.19. Or on a vu qu'il était impossible de faire passer le nombre d'enroulement d'un lacet de 0 au degré de P sans passer par le point 0 : c'est donc que notre hypothèse d'absence de racine pour P était absurde ; Cqfd.

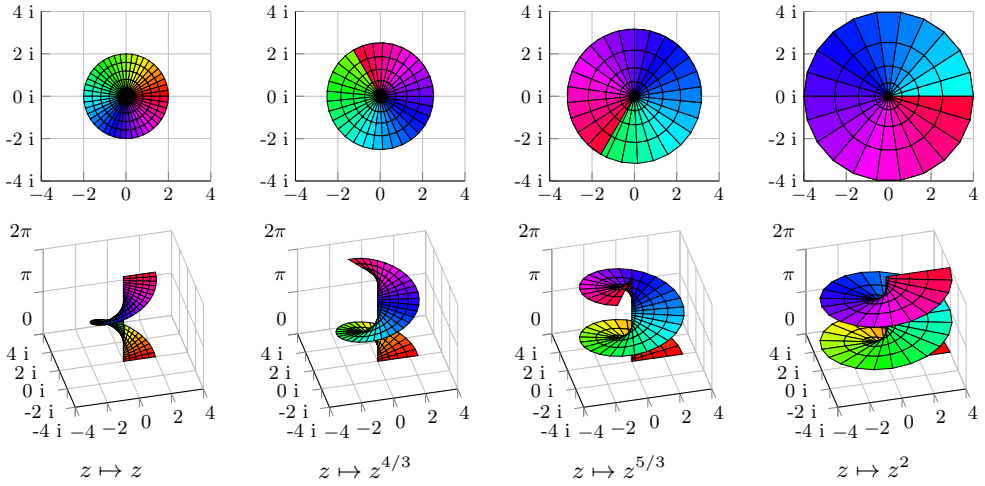
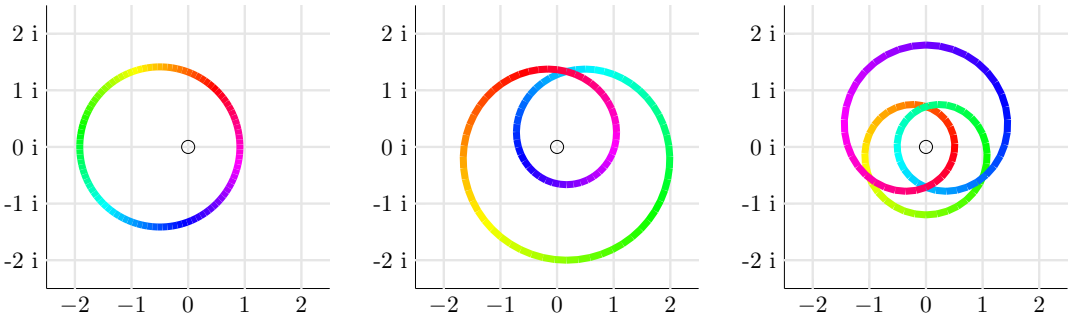


FIGURE 2.17 – Passage de l'identité $z \mapsto z$ au polynôme $z \mapsto z^2$, illustré sur le disque de centre 0 et de rayon 2. Pour aider à la visualisation, une vue éclatée est proposée dans la ligne du bas : l'image du complexe $\rho e^{i\theta}$ par l'application $z \mapsto z^n$ est représentée par le point $(\rho^n \cos(n\theta), \rho^n \sin(n\theta), \theta)$ On voit donc que chaque complexe non nul possède exactement 2 antécédents par l'application $z \mapsto z^2$, ce qui n'était pas évident sur la ligne du haut.



(a) $P(z) = (1+i)z - \frac{1}{2}$, $R = 1$. (b) $P(z) = \frac{i}{3}z^2 - \frac{1}{3}z + \frac{1}{6}$, $R = 2$. (c) $P(z) = z^3 - \frac{i}{5}z^2 - \frac{1}{2}z + \frac{i}{10}$, $R = 1$.

FIGURE 2.18 – Exemples de lacets images sous l'action de polynômes complexes.

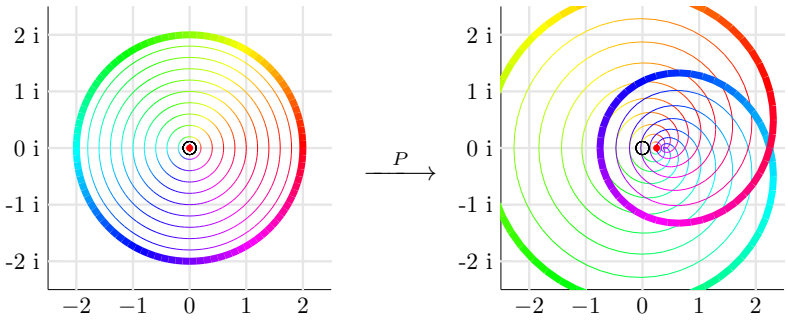


FIGURE 2.19 – Homotopie l associée au polynôme $P(z) = \frac{1}{2}z^2 + \frac{i}{2}z + \frac{1}{4}$ pour $R = 2$. À mesure que le rayon r croît de 0 à R, le lacet image l_r se déploie, passant d'un simple point $P(0)$ – en rouge – au bel enroulement l_2. Le nombre d'enroulements passant de 0 à 2, le théorème 2.2 affirme que notre cinématique a nécessairement survolé le point 0.

Preuve par relèvement

Applications conformes Les preuves précédentes ont mis en avant deux propriétés des polynômes complexes : ils sont localement surjectifs ; ils font tourner le disque à l'infini. La preuve suivante vont maintenant insister sur un point remarquable : un polynôme complexe est *conforme*, ou *holomorphe*, dérivable au sens complexe. Dans le film *Dimensions*, Adrien Douady définissait cette propriété par la *conservation des formes*, des angles. Un petit bouton de chemise restait envoyé sur un disque... Plus formellement :

Définition 2.1 (Application holomorphe)**.** Une application f de $\mathbb{C}$ dans $\mathbb{C}$ est dite *holomorphe*, ou *dérivable au sens complexe*, s'il existe une application dérivée f' de $\mathbb{C}$ dans $\mathbb{C}$ telle que :

$$\forall\ z \in \mathbb{C},\quad \frac{f(z+h)-f(z)}{h} \xrightarrow{h\to 0} f'(z), \tag{2.41}$$

ce que l'on préférera souvent noter

$$\forall\ z \in \mathbb{C},\quad f(z+h) =_0 f(z) + f'(z)\cdot h + o(h). \tag{2.42}$$

En tout point z, f est donc localement, à l'ordre 1, donnée par une similitude composée de la translation de vecteur $f(z)$, de l'homothétie de rapport $|f'(z)|$ et de la rotation d'angle $\mathrm{Arg}(f'(z))$.

On a alors le résultat suivant :

Proposition 2.1 (Les applications polynomiales sont holomorphes)**.** *Soit $P : z \mapsto \sum_{k=0}^{d} p_k z^k$ une application polynomiale de degré d. Alors P est holomorphe, et sa dérivée complexe est donnée par le polynôme de degré $d-1$,*

$$P' : z \mapsto \sum_{k=0}^{d-1} (k+1) p_{k+1} z^k. \tag{2.43}$$

Démonstration. Par linéarité de la limite, il suffit de le montrer pour les applications monomiales de la forme $z \mapsto z^d$. Or on a (formule de Pascal) :

$$\forall\ z \in \mathbb{C},\ \forall\ h \in \mathbb{C},\quad (z+h)^d = \sum_{k=0}^{d} \binom{d}{k} z^{d-k} h^k \tag{2.44}$$

$$= z^d + d\, z^{d-1}\, h + h^2 \cdot \Big(\cdots\Big), \tag{2.45}$$

ce qui se comprend très bien : en développant le produit de d termes "$(z+h)$", on fera apparaître un terme "z^d" (qui correspond au choix du z dans chacune des parenthèses), d termes "$z^{d-1}\, h$" (qui correspondent aux d choix possibles du type "je prends h dans une parenthèse, et z dans les autres"), ainsi que des termes d'ordre 2 et plus en h (qui correspondent aux choix ou l'on prend h dans plus d'une parenthèse). Résultat :

$$\frac{(z+h)^d - z^d}{h} = \frac{z^d + d\, z^{d-1}\, h + h^2 \cdot \big(\cdots\big) - z^d}{h} = d\, z^{d-1} + h \cdot \big(\cdots\big) \xrightarrow{h\to 0} d\, z^{d-1}. \tag{2.46}$$

□

Un polynôme non-constant n'a qu'un nombre fini de point critiques La dérivé d'un polynôme P de degré d est donc donnée par un polynôme P' de degré $d-1$.

Si P était constant, alors $P'=0$.

Sinon, on sait que P' ne peut avoir plus de $d-1$ racines : on peut en effet montrer, par un argument de division euclidienne, qu'un polynôme Q de racines $z_1, \ldots, z_k$ peut être factorisé par le polynôme $(z-z_1)\cdots(z-z_k)$, ce qui est impossible dès que k dépasse le degré de Q, à moins que Q soit nul.

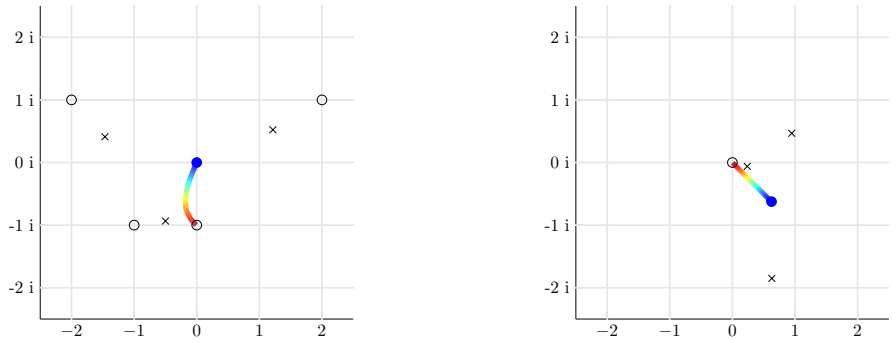
Ainsi, en dehors d'un nombre fini de points (qu'on appelle les *points critiques* de P), notre polynôme P non-constant est localement donné par une similitude *inversible* :

$$\text{à} \quad h \mapsto P(z) + P'(z)\cdot h \qquad \text{on peut opposer} \qquad s \mapsto \frac{1}{P'(z)}(s - P(z))\,. \tag{2.47}$$

Moralement, il est donc possible d'inverser localement le polynôme P au voisinage de $P(z)$, à condition que z ne soit pas un point critique de P – il faut que l'on puisse diviser par $P'(z)$.

Preuve du théorème fondamental de l'algèbre Soit donc P un polynôme non constant : il s'agit de trouver une racine z de P telle que $P(z)=0$. L'idée de la preuve *par relèvement*, illustrée Figure 2.20 est la suivante. Nous savons que P n'a qu'un nombre fini de point critiques : il est donc possible (facile !) de prendre z_0 qui ne soit pas un tel point critique. Dans l'espace "image", les images des points critiques, ou *valeurs critiques*, sont toujours en nombre fini. Alors, de deux choses l'une : si 0 est une valeur critique, c'est qu'il a un antécédent par P... Sinon, c'est qu'on peut trouver un chemin lisse $t \in [0,1] \mapsto \gamma_t$ qui évite toutes les valeurs critiques, et tel que $\gamma_0 = P(z_0)$, $\gamma_1 = 0$. Le long de ce chemin, P est **localement inversible**, et le facteur de dilatation en "$1/|P' \circ P^{-1}(\gamma_t)|$", continu sur le segment $[0,1]$, est uniformément borné.

"De proche en proche" – en réalité, via la théorie des équations différentielles et le théorème de Cauchy-Lipschitz, section B.4.2 –, on peut donc inverser P le long du chemin γ_t, et obtenir un chemin z_t tel que pour tout instant, $P(z_t) = \gamma_t$. À l'arrivée, z_1 est nécessairement une racine de P ; Cqfd.



(a) Espace des antécédents. Les cercles blanc représentent les racines de P ; les croix noires, les racines de P', ou *points critiques* de P.

(b) Espace des images. Les images des points critiques, ou *valeurs critiques* sont représentées par des croix noires ; le disque bleu désigne ici l'image par P du point 0.

FIGURE 2.20 – Relèvement d'un chemin de l'espace image vers celui des antécédents pour le polynôme $P(z) = \frac{1}{8}z^4 + \frac{1}{8}z^3 - \frac{2+i}{8}z^2 - \frac{3+8i}{8}z + \frac{5-5i}{8}$. Dans l'espace image (à droite), il est possible de trouver un chemin continu γ_t entre $P(z_0)$ et 0 qui évite les quelques valeurs critiques de P – segment arc-en-ciel.

Conclusion : la conception mathématique de la vérité

Que retenir de la première partie de ce cours ? Avant tout, la *relativité* des énoncés mathématiques. C'est une surprise : sous l'influence de Descartes, on se présente souvent les mathématiciens comme les platoniciens de la science moderne, amoureux passionnés de "La Vérité" qui transcende seule l'expérience sensible... S'il fut (partiellement) vrai jusqu'à l'époque des Russel, Hilbert ou Cantor, ce cliché a pris un bon coup de vieux au tournant du XX^{e} siècle.

On sait aujourd'hui, grâce à Gödel, qu'une théorie raisonnable des mathématiques ne saurait démontrer sa propre cohérence et ne peut donc être "auto-suffisante". Plus important encore, les grandes théories "historiques" que sont la géométrie d'Euclide, l'arithmétique des entiers, la combinatoire ou l'analyse réelle ne sont maintenant plus comprises que comme des cas particuliers, des cadres confortable qui n'ont rien d'*absolu*. Les géométries Riemanniennes, les algèbres abstraites, la théorie des ensembles transfinis ou la théorie des corps p-adiques sont passés par là : d'autres théories existent, qui sont au moins aussi naturelles que celles des anciens. (En un certain sens, elles le sont même plus : voir les fondements mathématiques de la physique relativiste, de la mécanique quantique !)

Alors, quelle valeur accorder au travail des mathématiciens ? À l'aube du XXI^{e} siècle, le temps des controverses philosophiques semble bien révolu. Rares sont maintenant ceux qui prétendent apporter une contribution d'ordre métaphysique au travers d'une recherche sur les nombres entiers ou les décimales de π... Par son travail, le mathématicien moderne cherchera avant tout à proposer un point de vue "*qui fait sens*" sur une question "*pertinente*".

Sans échelle de valeur absolue (une preuve de l'existence de Dieu par $e^{i\pi} + 1 = 0$ étant maintenant jugée inaccessible), les critères de *pertinence* d'une question varient d'un domaine à l'autre. En mathématiques *appliquées*, il suffira d'éclairer un problème *concret* ; de fournir aux ingénieurs des méthodes robustes, performantes et bien comprises. En mathématiques *fondamentales*, faire le consensus peut paraître plus difficile à mesure que s'éloignent les retombées pratiques... À la croisée de multiples domaines et spécialités, les "grandes conjectures" permettent aux mathématiciens de mesurer leurs progrès sur des échelles communes.

On a déjà parlé de la résolution des équations de Navier-Stokes (en analyse fonctionnelle), de l'hypothèse du continu (en théorie des ensembles) ou de la conjecture de Syracuse (en arithmétique), mais il ne s'agit là que des plus connues : chaque sous-domaine possède une liste de "questions ouvertes", étudiées depuis de longues années par des spécialistes passionnés. Au delà de leurs intérêts propres, les grandes questions fascinent par leurs histoires, les difficultés qui leur sont associées – on pense par exemple à la conjecture de Fermat. Les plus fameuses se remarquent par les connexions, les liens profonds qu'elles forcent à tisser entre des domaines a priori étrangers. On a discuté des multiples preuves du théorème fondamental de l'algèbre, des milles façons de parler des polynômes complexes. Tout en haut de l'échelle se trouve l'*Hypothèse de Riemann*, ouverte depuis 1859 : avec ses centaines de reformulations, c'est sans conteste l'Everest des mathématiques modernes.

La simple vérification de théorèmes "vrais" n'est donc pas une fin en soi. Ici comme ailleurs, l'arbitre des élégances restera le comité d'attribution des postes universitaires, INRIA, et CNRS : exigence d'excellence scientifique jugée par les pairs, mais aussi adéquation des thèmes de recherche avec les grands enjeux du moment. Préservation du rayonnement de l'école française d'algèbre, soutien à l'industrie aéronautique par un effort appuyé en mécanique des fluides, lancement du "plan Alzheimer" ou positionnement dans le monde des Big Data... Autant de questions qui orientent les intérêts, et façonnent la pensée mathématique de demain.

Chapitre 3

Analyse de Fourier : l'ubiquité d'une représentation

Séances 3 et 4

Au chapitre précédent, nous avons appris à penser les fonctions non plus comme des *graphes*, mais comme des *vecteurs* en dimension infinie. Chaque valeur $f(x)$ devenant une *coordonnée* associée à un point x, on peut bien penser à munir les espaces de fonctions de *repères*, orthonormés ou non, avec une infinité de directions. Rappelons-le : en dimension finie, on peut écrire un vecteur $x \in \mathbb{R}^d$ sous la forme

$$x \;=\; \sum_{i=1}^{d} x_i \cdot e_i, \qquad \text{avec une somme discrète,} \tag{3.1}$$

où e_i est le vecteur dont toutes les coordonnées sont nulles, sauf la i^{e} qui vaut 1. On dit que $(e_i)_{i\in[\![1,d]\!]}$ est la *base canonique* de $\mathbb{R}^d$, et que les x_i sont les coordonnées de x dans celle-ci. Eh bien, de manière analogue, si f est une fonction de $\mathbb{R}$ dans $\mathbb{R}$, on pourra dire que

$$f \;=\; \int_{-\infty}^{+\infty} f(x) \cdot \delta_x \,\mathrm{d}x, \qquad \text{avec une somme continue,} \tag{3.2}$$

au sens où pour toute fonction test φ, on a bien

$$\langle f, \varphi \rangle \;=\; \int_{-\infty}^{+\infty} f(x) \cdot \varphi(x) \,\mathrm{d}x \;=\; \int_{-\infty}^{+\infty} f(x) \cdot \langle \delta_x, \varphi \rangle \,\mathrm{d}x \;=\; \left\langle \int_{-\infty}^{+\infty} f(x) \cdot \delta_x \,\mathrm{d}x, \varphi \right\rangle, \tag{3.3}$$

par linéarité du crochet de dualité – pour peu que f vérifie quelques hypothèses techniques. En un sens, la famille des *diracs* $(\delta_x)_{x\in\mathbb{R}}$ forme donc la *base canonique* de l'espace des fonctions définies sur $\mathbb{R}$. Représenter f comme la donnée de ses valeurs $f(x)$ c'est donc, simplement, expliciter ses coordonnées dans la base canonique. Eh bien maintenant, pourquoi ne pas **changer de base** ?

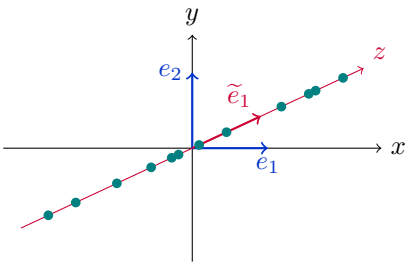


FIGURE 3.1 – Jeu de données dans le plan, un ensemble de points (x_i, y_i) dans $\mathbb{R}^2$ – en vert. Se focaliser sur ces coordonnées, c'est manquer l'évidence : un paramètre réel z associé au vecteur $\widetilde{e}_1$ suffit à décrire complètement la distribution de points.

Un problème pratique : la compression d'images

Commencer par choisir une représentation adaptée à son problème, c'est la première bonne idée du mathématicien. Nous avons vu dans le chapitre introductif du cours comment répondre au « problème de Loyd » de manière efficace, par un simple réarrangement des tuiles du jeu de taquin. Le problème d'aujourd'hui sera autrement plus pertinent, puisqu'il structure l'essentiel des communications sur le web : **comment transmet-on efficacement une *image* ?**

Préliminaires : l'encodage naïf des images

Un appareil photo numérique produit des *bitmaps*, tableaux numériques bidimensionnels qui encodent une image comme une grille de *pixels* monochromatiques et carrés. À l'état brut, une photo numérique « 3648x2736 » n'est rien d'autre que la donnée de $3648 \times 2736 = 9\,980\,928$ tuiles colorées, elles-mêmes encodées sur nos machines par un entier (image en noir et blanc) ou un triplet d'entiers (en couleurs, système RGB). Chacun de ces entiers étant (typiquement) encodé sur 8 bits, on aura alors accès à $2^8 = 256$ nuances de gris (entre le noir, 0 et 255, le blanc) et $2^8 \cdot 2^8 \cdot 2^8 = 16\,777\,216$ couleurs RGB par pixels.

Abstraction faite de certains détails liés à la structure physique des capteurs, on peut donc penser une image numérique comme un *vecteur* à n coordonnées entières, où n est de l'ordre du million.

Compression

On l'aura compris, un appareil photo génère un flot de données conséquent : près de $3 \times 3648 \times 2736 \simeq 30$ Mégaoctets pour une simple photo souvenir. Sur un réseau 3G à 384 ko/s (en zone urbaine), transmettre la moindre photo au format *bitmap* brut prend donc plus d'une minute. Or on le sait bien, une telle connexion permet "normalement" de regarder des vidéos en direct, sans altération notable de la qualité... C'est qu'un algorithme particulièrement efficace travaille en coulisses !

Pour libérer toutes les potentialités du réseau internet, le rôle du mathématicien est essentiel : il aura à *comprimer* les fichiers images de manière efficace, c'est à dire à les *réduire* en des fichiers de petites tailles, peu redondants, qui contiendront l'essentiel de l'information utile. Le résultat obtenu sera à l'image originale ce que le jus d'orange déshydraté est au Tropicana : un succédané sans finesse, mais bien moins encombrant.

Dans ce cours d'introduction, nous illustrerons nos idées sur l'image de référence la plus célèbre du *signal processing*, excellent exemple d'image dite "naturelle" : *Lena*, une photo 256x256 en niveaux de gris qui présente un visage lisse et de beaux dégradés, un plumeau particulièrement travaillé et de nombreuses plages texturées.

Une méthode naïve, la quantification Une fois la photo prise, comment en réduire le poids sur notre disque dur ? À l'état brut, il s'agit rappelons-le d'une liste de $256^2 \simeq 65\,000$ entiers encodés comme des nombres binaires à 8 chiffres :

indice	Écriture Binaire								Décimale	Couleur
	128	64	32	16	8	4	2	1		
1	0	0	0	0	0	0	0	0	0	
2	0	1	1	0	0	1	0	1	101	
3	1	1	1	1	1	1	1	1	255	
4	1	0	0	0	0	1	0	0	132	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
65536	0	0	0	1	0	1	1	1	23	

En mémoire, l'image est encodée par ce gros tableau binaire de taille 65536×8. Supprimer des lignes revient à oublier la couleur de certains pixels, et donc à rogner certaines parties de l'image ou à sous-échantillonner notre fichier. C'est, bien sûr, une possibilité présentée Figure 3.2.

De manière plus intéressante, on peut essayer de supprimer les colonnes de droite du tableau, qui encodent les bits de poids *faibles*. Ce procédé qu'on appelle *quantification* est illustré Figure 3.3 : il offre de piteuses performances.

FIGURE 3.2 – Effets du sous-échantillonnage sur le visage de Lena. À gauche, l'image originale, progressivement dégradée par l'oubli, sur le tableau *bitmap* de 3 lignes sur 4, 7 lignes sur 8 puis 15 lignes sur 16. La compression est brutale, avec un artefact de *blocking* immédiatement perceptible. Il est heureusement possible d'être plus efficace !
Image tirée du site de Ruye Wang, `fourier.eng.hmc.edu/e161/lectures/digital_image/node3.html`.

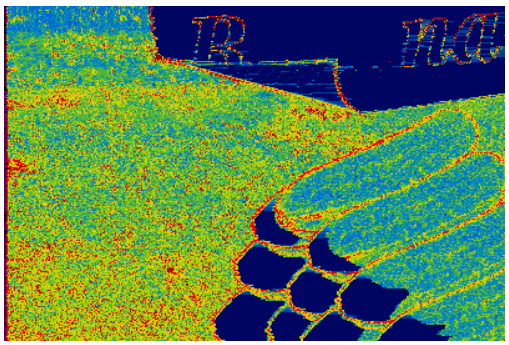
FIGURE 3.3 – Effets de la quantification binaire sur le visage de Lena. De gauche à droite et de haut en bas, on observe la même image de base encodée sur 256, 128, 64, 32, 16, 8, 4 et 2 niveaux de gris. Ces images correspondent respectivement à des indices de quantification de 1, 2, 4, 8, 16, 32, 64 et 128, et sont obtenues en "oubliant" 0, 1, 2, 3, 4, 5, 6 ou 7 des 8 colonnes à droite du tableau *bitmap*. Les niveaux de compression sont médiocres (1, 7/8, 6/8, 5/8, 4/8, 3/8, 2/8 et 1/8), au prix d'une dégradation considérable de la qualité de l'image : tous les dégradés sont remplacés par des aplats, et aucun a priori sur la structure de l'image n'est mis à profit.
Image tirée du site de Ruye Wang, `fourier.eng.hmc.edu/e161/lectures/digital_image/node2.html`.

Format PNG et codages entropiques Pour aller plus loin, on peut chercher à repérer les séquences récurrentes dans notre tableau *bitmap*. Ainsi, si le code 00010111 (couleur 23/255, gris foncé) revient très souvent, on peut essayer de lui associer un code plus court (disons, 110) quitte à donner un code un plus long à une couleur rarement utilisée. À la limite, on peut même appliquer cette idée à des blocs de pixels contigus récurrents, par exemple des aplats de couleur uniformes.

Trouver un encodage optimal pour une suite numérique arbitraire relève plus de l'informatique (la science de l'*information*) que des mathématiques. Nous ne détaillerons donc pas ici les algorithmes optimaux, qui reposent sur la théorie des encodages entropiques dont les plus célèbres sont le *code de Huffman* voire l'algorithme *Lempel–Ziv–Welch*. Utilisés par le format d'archivage *.zip* comme par le format d'images *.png*, cette théorie fournit des algorithmes de codage *sans perte*, qui peuvent être optimaux au sens où ils sont asymptotiquement plus performants que tout code reposant sur une table de correspondance (ou *dictionnaire*) finie.

Illustrées Figure 3.4, les performances de cet algorithme sont satisfaisantes sur les images simples, les dessins, les scans de bandes dessinées ou de textes... Mais, sans critères bien posés de perte acceptable ou de régularité, il se retrouve à la peine pour comprimer des images naturelles comme des photos, où les séquences peuvent se ressembler sans être jamais identiques d'un endroit à un autre. Sur internet vont donc coexister deux formats d'images dominants : le *.png*, qui repose sur la théorie de l'*information* et est particulièrement adapté aux dessins et schémas ; le *.jpg*, adapté spécifiquement à la compression de photos naturelles, que nous allons maintenant présenter. En utilisant à fond un a priori de régularité sur l'image à comprimer, il fournira un parfait exemple d'application des idées de la géométrie euclidienne aux problèmes industriels modernes.

(a) Image originale présentant un mélange d'aplats, de dégradés, de détails fins et de régions texturées.



(b) Coût en bit de chaque pixel : les zones colorées en rouge sont chères, le bleu correspond aux régions comprimées de manière efficace.

FIGURE 3.4 – Illustration de l'algorithme de compression utilisé par le format *.png*, qui repose sur un pré-traitement ad hoc et un codage de Huffman ligne par ligne. Les aplats uniformes sont très fortement comprimés, tandis que l'essentiel du coût se concentre dans les régions détaillées de l'image, les petits villages. On remarquera que la pile de bananes – obtenue par "copier-coller" d'une unique image – est ici parfaitement interprétée : si la banane du haut est encodée avec difficultés, les extrémités suivantes sont identifiées comme étant de simples copies et les codes mis au point pour la première banane sont réutilisés.
Image tirée de Wikipédia, par Pink kitty111.

Le format JPEG, un simple changement de repère

Exception faite du pré-traitement effectué par le format PNG, les méthodes précédentes se contentaient de voir le tableau *bitmap* comme un flot de données arbitraire, à comprimer par un algorithme générique. Le sous-échantillonnage, la quantification, l'encodage de Huffman... Toutes ces méthodes peuvent s'appliquer à un fichier son *.wav* comme à une image *.bmp*.

Or on le sait bien : une photo et une chanson, ce n'est pas la même chose ! Pour mettre au point des algorithmes *vraiment* efficaces, il faudra tirer au mieux parti de la structure de nos jeux de données.

Pixels et base canonique Penser les images comme une juxtaposition de *pixels* est devenu naturel aux enfants des années 90 que nous sommes : MS Paint, Mario et le pixel art sont passés par là ! Mais qu'est-ce au juste qu'un tableau *bitmap* ? Mathématiquement, rien d'autre que l'écriture de notre vecteur "image" dans la *base canonique* associée à la grille d'échantillonnage. Pour $(i,j) \in I \times J = [\![0,255]\!] \times [\![0,255]\!]$ un indice arbitraire, on peut en effet définir l'image dite "de Dirac" $\delta_{i,j}$ par :

$$\delta_{i,j}(u,v) \;=\; \begin{cases} 1 & \text{si } u=i \text{ et } v=j \\ 0 & \text{sinon} \end{cases}. \tag{3.4}$$

C'est l'analogue bidimensionnel des vecteurs e_i de la base canonique, utilisés dans l'équation (3.1). Une image x quelconque s'écrira alors :

$$x \;=\; \sum_{i,j} x_{i,j} \cdot \delta_{i,j}, \tag{3.5}$$

où $x_{i,j}$ est la "valeur" du pixel de x de coordonnées (i,j).

Adopter ce point de vue algébrique, c'est penser la décomposition pixel à pixel non plus comme une *juxtaposition*, mais comme une *superposition* de calques très localisés – les $\delta_{i,j}$ – dont les intensités sont modulées par les coefficients $x_{i,j}$ – voir Figure 3.6. Par là, une direction de recherche s'ouvre tout à coup : pourquoi ne pas remplacer les $\delta_{i,j}$ par une nouvelle famille ?

Une base de calques adaptée aux images naturelles Pour comprimer nos données, nous sommes à la recherche d'un système d'écriture *compact*. On peut comprendre cette problématique par une petite analogie postale : si chaque octet d'information transmis correspondait à un calque d'un gramme, transmettre les 65 536 pixels de l'image Lena reviendrait à envoyer un classeur de transparents par la poste... pour un poids total de 65kg et une addition salée.

Or, en regardant de plus près nos transparents $x_{i,j} \cdot \delta_{i,j}$, on remarquerait que ceux-ci sont vides presque partout avec une seule position colorée (i,j). Quel gâchis ! Au cœur de la compression JPEG, se trouve une idée fondamentale : l'utilisation d'une nouvelle famille de calques de base, les $f_{i,j}$ présentés Figure 3.6 et 3.7, définis sur des blocs 8x8 par

$$f_{i,j}(u,v) \;=\; \cos\left(\frac{i\pi}{8}\left(u+\frac{1}{2}\right)\right) \cdot \cos\left(\frac{j\pi}{8}\left(v+\frac{1}{2}\right)\right) \tag{3.6}$$

pour i, j, u et v dans $[\![0,7]\!]$ – on commence à compter les indices à partir de 0, pour des raisons pratiques. Par définition, $f_{i,j}$ sera donc une image 8x8 à valeurs dans $[-1,1]$ présentant i rayures sur la première direction et j sur la seconde. Après un calcul qui peut être effectué à moindre coût sur des puces électroniques dédiées, on obtient pour toute image x une décomposition sur la base $f_{i,j}$ en coefficients $c_{i,j}$, et on écrira :

$$x \;=\; \sum_{i,j} x_{i,j} \cdot \delta_{i,j} \;=\; \sum_{i,j} c_{i,j} \cdot f_{i,j}. \tag{3.7}$$

Parcimonie Un constat s'impose : pour reformer la joue ou l'épaule de Lena, les grandes rayures suffisent... et il n'est guère utile d'avoir à disposition des calques de damiers ! Si l'image x est régulière, tirée d'une image naturelle, alors la transformée $c_{i,j}$ présente un profil dont nous allons pouvoir tirer parti : les coefficients $c_{i,j}$ correspondant à des nombres de rayures élevés (i ou j supérieurs à 3) sont très petits.

En première approximation, on peut donc se contenter d'envoyer par la poste les 6, 10 ou 15 coefficients $c_{i,j}$ correspondant aux petites valeurs de $i + j$, comme illustré Figure 3.10. En n'envoyant que les 10 calques "les plus importants" pour chaque bloc de 8x8 pixels, on diviserait par 6 le poids du colis en altérant peu le rendu final. La solution finalement retenue par les chercheurs est encore plus astucieuse : en utilisant un tableau de *quantification* fixé à l'avance – voir Figure 3.8 –, on restreint le nombre de bits alloués aux fortes valeurs de $i + j$ sans les éliminer tout à fait, ce qui permet une transition moins brutale entre coefficients "conservés" et coefficients "oubliés".

Tous comptes faits, la décomposition dans une base de cosinus *adaptée* aux images naturelles permet à notre algorithme de filtrer simplement les composantes peu utiles de notre image, celles qui correspondent aux motifs en damier hautes fréquences. Nettement plus adapté aux données réelles qu'un algorithme de sous-échantillonnage naïf, le format JPEG sera donc très performant sur les photos naturelles. Simple à comprendre et à implémenter, il aura le succès que l'on connaît. Sur les figures qui suivent, je vous propose de découvrir pas à pas les détails d'un algorithme qui est au cœur de l'imagerie numérique grand public.

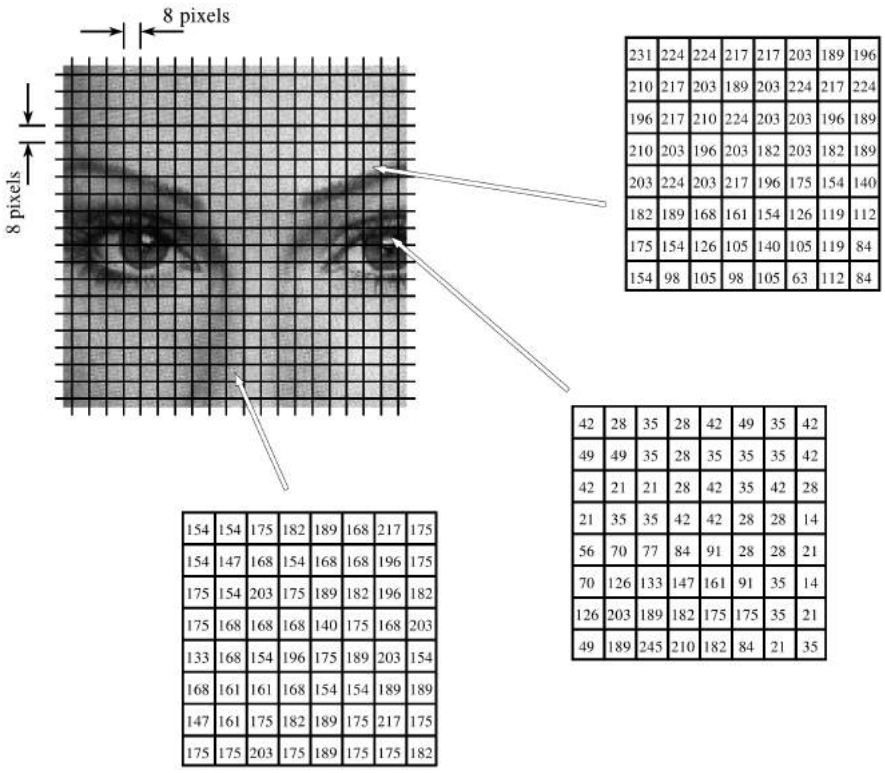
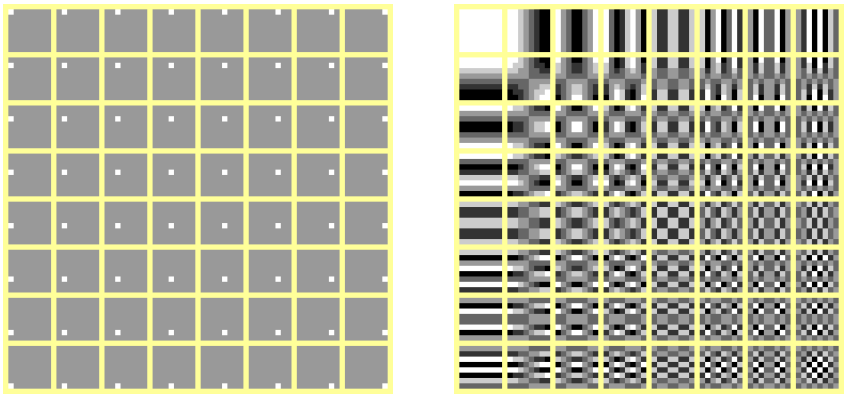


FIGURE 3.5 – Première étape de la compression JPEG : une découpe en blocs de 8x8 pixels qui seront traités indépendamment – modulo la couleur moyenne.
Image tirée du site web `www.dspguide.com`.



(a) Famille des diracs $\delta_{i,j}$. (b) Famille des cosinus discrets $f_{i,j}$.

FIGURE 3.6 – Représentation synthétique des deux familles utilisées pour l'encodage de blocs d'images 8x8 : à droite, la famille des diracs ou *base canonique* utilisée par les capteurs et format des données *bitmap* brutes ; à gauche, la famille des cosinus discrets employée par le format JPEG. Affichés sur fond jaune, ces deux groupes de 8×8 images de taille 8x8 forment des bases de l'espace $\mathbb{R}^{8\times 8}$ des images 8x8 : ici, le blanc correspond à une coordonnée de 1, le noir au -1 et le gris 128/255 au 0.
En admettant que ce dernier correspond à la "transparence", une image 8x8 peut être vue comme la *superposition* de 64 filtres dont les intensités sont modulées par des coefficients notés $x_{i,j}$ dans le cas de la base des diracs, et $c_{i,j}$ dans le cas de la base des cosinus discrets.
La donnée des coefficients $x_{i,j}$, c'est un tableau *bitmap* difficile à comprimer. Le bloc 8x8 des coefficients $c_{i,j}$ constitue quand à lui la *transformée en cosinus discrets* du bloc, qui sera nettement plus facile à manipuler que le tableau bitmap initial.

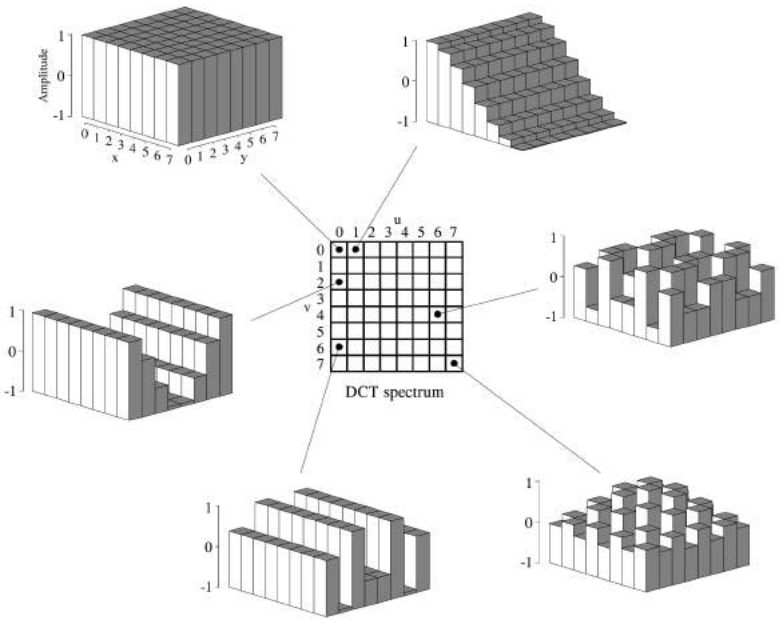
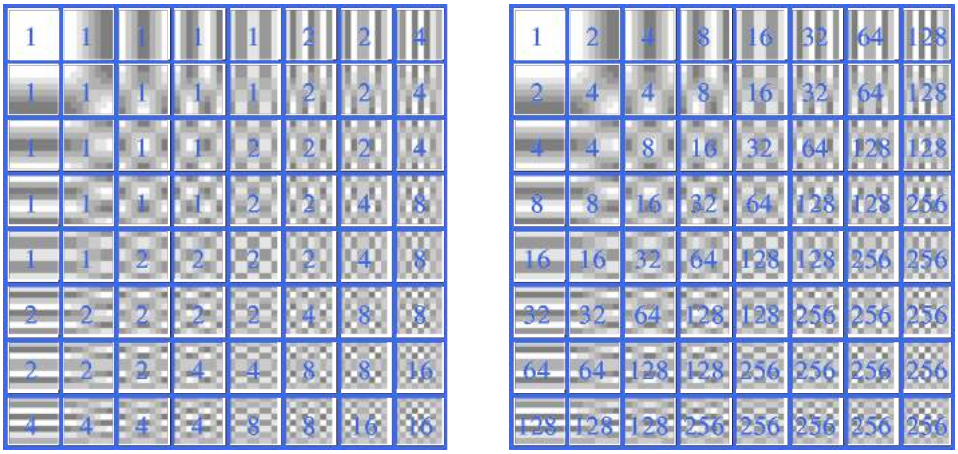


FIGURE 3.7 – Deuxième étape de la compression JPEG : la transformée en cosinus discrète sur l'espace des images de taille 8x8. Il s'agit d'un analogue immédiat de la transformée de Fourier décrite plus loin, optimisé pour le traitement de signaux *réels* non-périodiques. Un élément $f_{i,j}$ de la base est caractérisé par ses coefficients i et j entre 0 et 7, qui quantifient le nombre de ses oscillations : i dans la direction verticale, j en horizontale. D'une représentation point-par-point de notre image, on passe donc à une représentation fréquentielle, avec les basses fréquences stockées en haut à gauche de chaque "bloc".
Image tirée du site web `www.dspguide.com`.



(a) Faible compression. (b) Forte compression.

FIGURE 3.8 – Troisième étape de la compression JPEG : les coefficients de la transformée sont *quantifiés* par l'oubli d'un nombre arbitraire de bits de poids faible, comme dans la Figure 3.3. Ces indices de quantification dépendent de la fréquence et du niveau de compression choisie : deux exemples de tables sont donnés ici. Attention : il ne s'agit pas ici d'images 8x8, mais bien de tableaux de coefficients d'échantillonnage, qui indiquent le degré d'imprécision sur chaque fréquence (i, j) dans le fichier final. C'est la seule étape *destructive* de la compression JPEG.
On remarquera que la moyenne, ou fréquence $(0, 0)$, est toujours préservée – indice de quantification égal à 1. Par contre, en compression élevée, les hautes fréquences sont affectées à des indices de quantification élevés : 64, 128 voire 256, soit une conservation de 2, 1 ou 0 chiffres de l'écriture binaire des coefficients.
Image adaptée du site web `www.dspguide.com`.

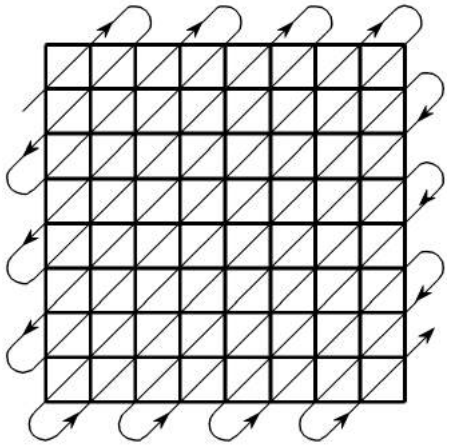


FIGURE 3.9 – Quatrième étape de la compression JPEG : on tire parti du grand nombre de zéros dans les hautes fréquences par un codage astucieux. Les coefficients de la transformée quantifiée sont d'abord mis en ligne, des basses aux hautes fréquences – le sens de lecture sur le bloc des indices (i, j) est donné par la figure ci-dessus. Une suite typique serait par exemple
. $(52, 32, 22, 4, 4, 8, 8, 8, 8, 8, 0, 0, 0, \dots)$.
On pourra alors la coder efficacement par un algorithme en "longueur de suite", qui la transformera par exemple en "1-52, 1-32, 1-22, 2-4, 5-8, 54-0", nouvelle suite compacte qui sera encodée (de manière efficace) sous forme de suite binaire : le fichier ".jpg" final.
Image tirée du site web `www.dspguide.com`.

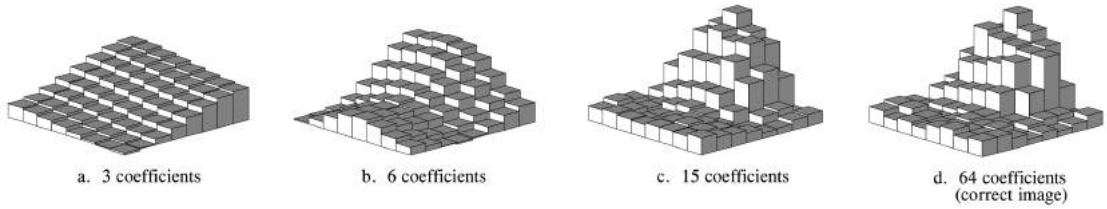
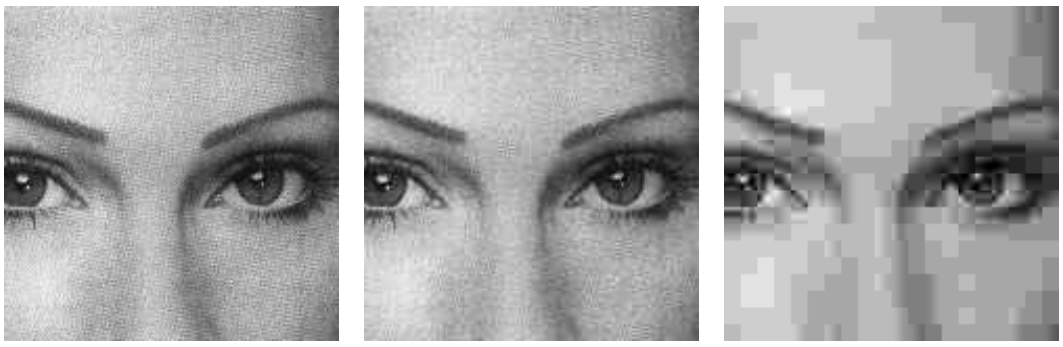
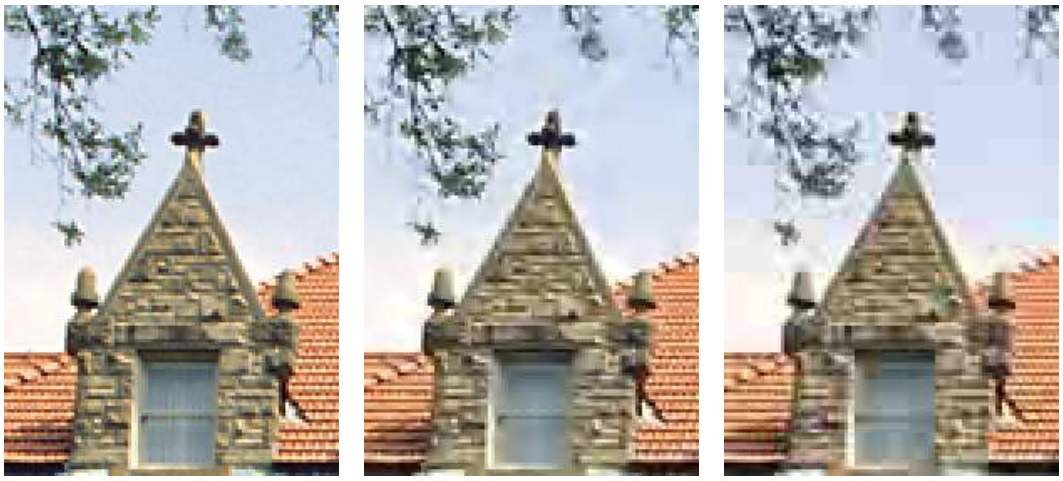


FIGURE 3.10 – Lecture du format JPEG, par reconstruction de chaque bloc d'image à l'aide des quelques coefficients stockés dans le fichier. Le bloc original se trouve à droite, et est vu ici "en 3D" – et non comme un tableau 8x8 de niveaux de gris ou de valeurs entières entre 0 et 255. Après transformée en cosinus, conservation des seuls n plus gros coefficients et reconstruction, on constate sans surprise une dégradation du signal. Heureusement, et c'est là tout l'intérêt du format JPEG, un résultat satisfaisant est tout de même obtenu avec une conservation de seulement 15 des 64 coefficients originaux : on est donc en droit d'espérer une perte en qualité négligeable pour des taux de compression de 4 : 1. Image tirée du site web `www.dspguide.com`.



(a) Image originale. (b) Compression de 10 : 1. (c) Compression de 45 : 1.

FIGURE 3.11 – Compression d'une photo de visage par l'algorithme JPEG. Celui-ci commence par découper l'image en blocs de 8x8 pixels, puis effectue sur chacun d'eux une transformée en cosinus ; il en élimine ensuite les petits coefficients. L'image étant généralement lisse sur chacun des blocs, le résultat net d'une forte compression est l'annihilation des hautes fréquences, phénomène particulièrement visible sur l'image de droite.



(a) Image originale. (b) JPEG2000, 20 : 1. (c) JPEG, 20 : 1.

FIGURE 3.12 – Comparaison de deux "générations" de l'algorithme de compression JPEG, avec dans les deux cas un ratio de compression de 20 : 1. L'algorithme JPEG2000 se veut le successeur du standard JPEG : à la simpliste transformée de Fourier par blocs a été substituée une transformée en ondelettes multi-échelle, qui élimine les artefacts de "blocking" et offre de nombreux avantages pratiques. Images tirées du site `www.photozone.de`.

Une base orthonormale pertinente

J'ai jusqu'ici fait le choix de vous initier à l'analyse harmonique par l'exemple, en vous montrant le parti que l'on pouvait tirer d'un **changement de base** astucieux. Nous allons maintenant décrire plus précisément les qualités que l'on attend d'une famille adaptée aux problèmes d'analyse : orthogonalité, diagonalisation de la dérivée. Ce sera l'occasion de découvrir le concept sous-jacent à la base des *cosinus discrets* : celui d'*harmoniques*.

Le produit scalaire, mesure de corrélation

Une base $(e_i)_{i\in[\![1,n]\!]}$ de $\mathbb{R}^n$ permet d'obtenir pour tout vecteur x une décomposition en coordonnées $(x_i)_{i\in[\![1,n]\!]}$. On souhaite typiquement que celles-ci soient aussi "indépendantes" les unes des autres que possible, décorrélées. De manière intuitive, on préférera donc une base dont les vecteurs sont *orthogonaux* les uns aux autres. Mais comment parler d'orthogonalité en dimension supérieure ou égale à quatre ?

Définition 3.1 (Produit scalaire)**.** Si $x = (x_1, \dots, x_n)$ et $y = (y_1, \dots, y_n)$ sont deux vecteurs de l'espace $\mathbb{R}^n$ de dimension n, on définit leur produit scalaire par

$$(\, x \,|\, y \,) \;=\; \sum_{i=1}^{n} x_i y_i \;=\; x_1 y_1 + \cdots + x_n y_n. \tag{3.8}$$

Le produit scalaire est une mesure réelle de la corrélation entre deux vecteurs, définie au travers de leurs coordonnées dans la base canonique. Il sera positif si les x_i et y_i sont souvent de même signe, négatif s'ils restent plutôt en opposition et nul si les deux suites de nombres sont totalement décorrélées. La positivité de $(\, x \,|\, y \,)$ est donc une indication du fait que "x et y pointent dans la même direction", ce qui est formalisé par la proposition suivante.

Proposition 3.1 (Inégalité de Cauchy-Schwarz, angle entre deux vecteurs)**.** *Soient $x = (x_1, \dots, x_n)$ et $y = (y_1, \dots, y_n)$ deux vecteurs de $\mathbb{R}^n$. On remarque que*

$$\|x\|_2 \;=\; \sqrt{(\, x \,|\, x \,)}, \tag{3.9}$$

où $\|\cdot\|_2$ est la « norme 2 » définie équation 10.3 et on a la domination dite de Cauchy-Schwarz :

$$|(\, x \,|\, y \,)| \;\leqslant\; \|x\|_2 \cdot \|y\|_2 \,. \tag{3.10}$$

Il existe alors un unique angle $\theta \in [0, \pi]$ tel que

$$(\, x \,|\, y \,) \;=\; \|x\|_2 \cdot \|y\|_2 \cdot \cos(\theta), \tag{3.11}$$

et on dira que θ est l'angle géométrique entre x et y. Le produit scalaire suffit donc à définir l'angle entre deux vecteurs quelconques de $\mathbb{R}^n$, à orientation près.

Démonstration. On définit le binôme

$$P(t) \;=\; \|t \cdot x + y\|_2^2 \;=\; (\, t \cdot x + y \,|\, t \cdot x + y \,) \;=\; t^2 \cdot (\, x \,|\, x \,) + 2\, t \cdot (\, x \,|\, y \,) + (\, y \,|\, y \,), \tag{3.12}$$

par bilinéarité de l'expression (3.8). Par définition, $P(t)$ est positif pour toute valeur réelle de t. Son discriminant

$$\Delta_P \;=\; 4\,(\, x \,|\, y \,)^2 - 4\,(\, x \,|\, x \,)\,(\, y \,|\, y \,) \tag{3.13}$$

est donc négatif ou nul, cqfd.

On remarquera que P s'annule dans $\mathbb{R}$ si et seulement si Δ_P est nul, c'est à dire si l'angle θ est égal à 0 ou π. Or l'existence d'une racine pour P équivaut à celle d'un réel t_0 tel que $y = -t_0 \cdot x$. Conformément à l'intuition, l'angle entre x et y est donc nul ou plat si et seulement si ces vecteurs sont colinéaires. □

Définition 3.2 (Orthogonalité, base orthonormale)**.** Si x et y sont deux vecteurs de l'espace euclidien $\mathbb{R}^n$ muni du produit scalaire canonique défini équation (3.8), on dira qu'ils sont orthogonaux entre eux si et seulement si $(\, x \,|\, y \,) = 0$.

Si $(e_i)_{i\in[\![1,n]\!]}$ est une famille de vecteurs de $\mathbb{R}^n$, on dira que c'est une *base orthonormale* de l'espace euclidien si et seulement si

$$\forall\, i,j \in [\![1,n]\!]\,,\ (\, e_i \,|\, e_j \,) \;=\; \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases}. \tag{3.14}$$

La base canonique utilisée dans équation (3.1) est un premier exemple important. La recherche de base orthonormales sera motivée par les propositions suivantes :

Proposition 3.2 (Base orthonormale duale)**.** *Soit x un vecteur de l'espace euclidien $\mathbb{R}^n$ muni d'une base orthonormale $(e_i)_{i\in[\![1,n]\!]}$. Alors les coefficients de x dans cette dernière se calculent au travers des produits scalaires $(\, e_i \,|\, x \,)$: on aura simplement*

$$x \;=\; \sum_i (\, e_i \,|\, x \,)\, e_i. \tag{3.15}$$

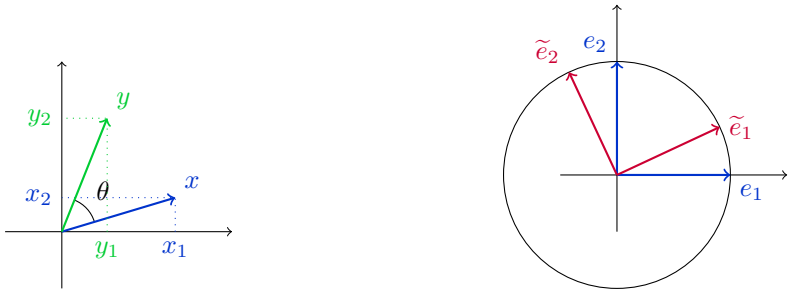
Démonstration. La famille des e_i possédant n vecteurs orthogonaux deux à deux, elle est *libre*. Par le théorème de la dimension (non démontré ici), il s'agit bien d'une base génératrice de l'espace $\mathbb{R}^n$, et on peut obtenir un n-uplet de coefficients x_i tels que

$$x \;=\; \sum_i x_i\, e_i. \tag{3.16}$$

Mais alors, pour tout indice j dans $[\![1,n]\!]$, on a

$$(\, x \,|\, e_j \,) \;=\; \sum_i (\, x_i\, e_i \,|\, e_j \,) \;=\; \sum_i x_i\, (\, e_i \,|\, e_j \,) \;=\; x_j. \tag{3.17}$$

On peut donc décomposer x dans la base des e_j par de simples calculs de produits scalaires. □



(a) Produit scalaire entre deux vecteurs x et y : on a $x_1y_1 + x_2y_2 = \|x\|\,\|y\|\,\cos(\theta)$.

(b) Exemples de bases orthonormales du plan $\mathbb{R}^2$, en bleu et en rouge.

FIGURE 3.13 – La géométrie des espaces *euclidiens* repose sur un produit scalaire directement calculable en coordonnées dans des bases dites *orthonormales* comme la base canonique.

Théorème 3.1 (Identité de Parseval)**.** *Soient x et y deux vecteurs de l'espace $\mathbb{R}^n$ muni d'une base orthonormale $(e_i)_{i\in[\![1,n]\!]}$. Alors on peut calculer le produit scalaire $(\,x\,|\,y\,)$ à partir des seules coordonnées de x et y dans la base des e_i :*

$$(\,x\,|\,y\,) \;=\; \sum_{i=1}^{n} (\,e_i\,|\,x\,)\,(\,e_i\,|\,y\,)\,. \tag{3.18}$$

En particulier, on a que :

$$\|x\|_2 \;=\; \sqrt{\sum_i (\,e_i\,|\,x\,)^2}. \tag{3.19}$$

Démonstration. Il suffit d'utiliser l'expression (3.15), puis de développer le double produit par bilinéarité du produit scalaire :

$$\begin{aligned}
(\,x\,|\,y\,) \;&=\; \left(\sum_i (\,e_i\,|\,x\,)\,e_i \,\middle|\, \sum_j (\,e_j\,|\,x\,)\,e_j \right) && (3.20)\\
&=\; \sum_{i,j} ((\,e_i\,|\,x\,)\,e_i \,|\, (\,e_j\,|\,x\,)\,e_j\,) && (3.21)\\
&=\; \sum_{i,j} (\,e_i\,|\,x\,)\,(\,e_j\,|\,x\,)\,(\,e_i\,|\,e_j\,) && (3.22)\\
&=\; \sum_i (\,e_i\,|\,x\,)\,(\,e_i\,|\,y\,)\,, && (3.23)
\end{aligned}$$

puisque $(\,e_i\,|\,e_j\,)$ est non nul si et seulement si $i = j$. □

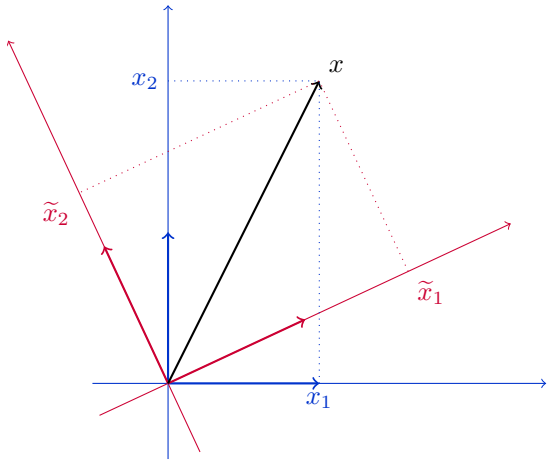


FIGURE 3.14 – L'identité de Parseval permet de calculer la norme d'un vecteur x dans n'importe quelle base orthonormale : on a ici $x_1^2 + x_2^2 = \|x\|_2^2 = \widetilde{x}_1^2 + \widetilde{x}_2^2$. C'est une conséquence directe de la bilinéarité du produit scalaire que l'on peut interpréter comme une application moderne du théorème de Pythagore.

Construction de bases orthonormales Les bases orthonormales sont donc celles qui sont aussi adaptées que la base canonique aux calculs de normes, de produits scalaires entre deux vecteurs. On a vu dans la première section du chapitre tout l'intérêt qu'il y avait à disposer d'une base comprenant le vecteur constant $\mathbf{1} = (1, \dots, 1)$, noté $f_{0,0}$ dans la base des cosinus discrets. En incarnant l'idée de valeur moyenne, il permet une analyse du signal délocalisée, plus sémantique que celle effectuée dans la base des diracs. Alors, **peut-on construire simplement une base orthonormale de $\mathbb{R}^n$ dont le premier vecteur soit un multiple du vecteur constant 1 ?**

Le cas des dimensions 2 et 4 Dans le plan euclidien, nous pouvons nous appuyer sur une solide intuition : il suffit de considérer la base formée des deux vecteurs

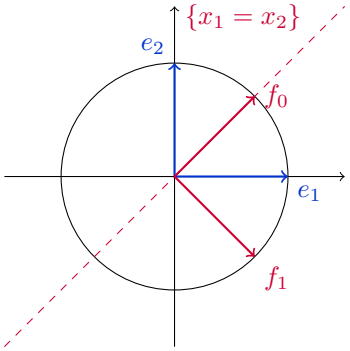
$$f_0 = \frac{1}{\sqrt{2}}(+1, +1) \qquad \text{et} \qquad f_1 = \frac{1}{\sqrt{2}}(+1, -1)\,, \tag{3.24}$$

qui est bien orthonormale. Par l'alternance des signes, f_1 parvient à être orthogonal à f_0 tout en restant de norme 1. On peut s'inspirer de cette première base pour définir en dimension $2 \times 2 = 4$ une famille analogue, obtenue par "produit tensoriel" :

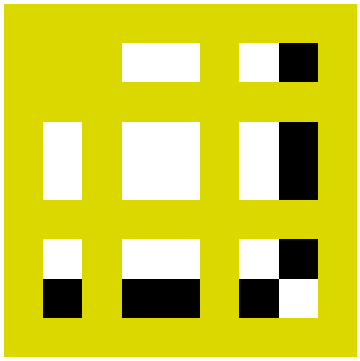
$$f_{0,0} = \frac{1}{\sqrt{4}} \begin{pmatrix} +1 & +1 \\ +1 & +1 \end{pmatrix}, \qquad f_{0,1} = \frac{1}{\sqrt{4}} \begin{pmatrix} +1 & -1 \\ +1 & -1 \end{pmatrix}, \tag{3.25}$$

$$f_{1,0} = \frac{1}{\sqrt{4}} \begin{pmatrix} +1 & +1 \\ -1 & -1 \end{pmatrix}, \qquad f_{1,1} = \frac{1}{\sqrt{4}} \begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}. \tag{3.26}$$

Il s'agit bien sûr de l'analogue de la famille des cosinus discrets pour les images 2x2... Mais comment généraliser cette construction aux dimensions quelconques ?



(a) Base (f_0, f_1) par rapport à la base canonique. Généraliser cette construction aux dimensions supérieures, c'est le défi relevé par les harmoniques discrètes.



(b) Base orthogonale des $(f_{0,0}, f_{0,1}, f_{1,0}, f_{1,1})$ obtenue par produit tensoriel des deux vecteurs f_0 et f_1. Ici, le blanc vaut $1/2$ et le noir $-1/2$; on présente les 4 images sur un fond jaune pour éviter toute confusion.

FIGURE 3.15 – En jouant sur les signes, on peut construire des bases oscillantes des espace de dimension 2^n.

Le grand retour des nombres complexes En restant prisonniers de la droite réelle, trouver des formules appropriées en toute dimension est extrêmement compliqué. C'est qu'il n'y a que deux nombres réels de norme 1 : $+1$ et -1. Les jeux d'alternances de signes ne pourront donc nous permettre de traiter que des dimensions qui sont puissances du nombre 2 comme 8, 16 ou 256.

Comment nous tirer de ce mauvais pas ? En nous plongeant dans le corps des complexes, où l'ensemble des nombres de modules 1 est infini : il s'agit du cercle unité

$$\mathbb{U} \;=\; \{z \in \mathbb{C}, |z| = 1\} \;=\; \left\{e^{i\theta}, \theta \in \mathbb{R}\right\}. \tag{3.27}$$

On peut définir sur $\mathbb{C}$ le produit scalaire canonique par

$$(\, z_1 \,|\, z_2 \,) \;=\; \mathrm{Re}(z_1) \cdot \mathrm{Re}(z_2) + \mathrm{Im}(z_1) \cdot \mathrm{Im}(z_2) \;=\; \mathrm{Re}(\overline{z_1} \cdot z_2), \tag{3.28}$$

puis sommer ces identités pour définir un produit scalaire sur $\mathbb{C}^n$ qui prolonge celui donné dans l'équation (3.8). La généralisation des équations (3.24-3.26) à l'espace vectoriel $\mathbb{C}^n$ est alors immédiate :

Définition 3.3 (Transformée de Fourier discrète)**.** On définit la famille des vecteurs harmoniques de taille n par :

$$\forall\; j \in [\![0, n-1]\!],\; f_j \;=\; \left(\frac{1}{\sqrt{n}}\, e^{\frac{2i\pi}{n} \cdot jk}\right)_{k \in [\![0,n-1]\!]} \;=\; \frac{1}{\sqrt{n}}\,(1, \omega^j, \omega^{j\cdot 2}, \omega^{j\cdot 3}, \ldots, \omega^{j\cdot(n-1)}) \tag{3.29}$$

où $\omega = e^{\frac{2i\pi}{n}}$ est un complexe de module 1 tel que $\omega^n = 1$. C'est le point du cercle qui délimite avec l'axe des abscisses un "n-ième" de disque, de sorte que les ω^k correspondent à une découpe équitable du disque en n parts.
Alors la famille $(f_j)_{j \in [\![0,n-1]\!]}$ forme une base orthonormale de $\mathbb{C}^n$, au sens où pour toute paire d'indices a et b, on a

$$(\, f_a \,|\, f_b \,) \;=\; \begin{cases} 1 & \text{si } a = b \\ 0 & \text{sinon} \end{cases}. \tag{3.30}$$

Démonstration. Pour a et b deux indices entiers dans $[\![0, n-1]\!]$, on a

$$(\, f_a \,|\, f_b \,) \;=\; \sum_{k=0}^{n-1} \overline{\left(\frac{1}{\sqrt{n}}\, e^{\frac{2i\pi}{n} \cdot ak}\right)} \cdot \left(\frac{1}{\sqrt{n}}\, e^{\frac{2i\pi}{n} \cdot bk}\right) \tag{3.31}$$

$$= \frac{1}{n} \sum_{k=0}^{n-1} e^{-\frac{2i\pi}{n} \cdot ak} \cdot e^{\frac{2i\pi}{n} \cdot bk} \tag{3.32}$$

$$= \frac{1}{n} \sum_{k=0}^{n-1} e^{\frac{2i\pi}{n} \cdot (b-a)k} \tag{3.33}$$

$$= \frac{1}{n} \sum_{k=0}^{n-1} \eta^k \tag{3.34}$$

où $\eta = e^{\frac{2i\pi}{n} \cdot (b-a)}$ est un complexe de module 1 tel que $\eta^n = 1$.

Si $b = a$, on a $\eta = 1$ puis $(\, f_a \,|\, f_b \,) = \frac{1}{n} \cdot n = 1$. Sinon, on a $0 < |b - a| < n$ puis $\eta \neq 1$. La formule de sommation des séries géométriques (que l'on démontre par récurrence) permet alors de trouver

$$(\, f_a \,|\, f_b \,) \;=\; \frac{1}{n} \frac{1 - \eta^n}{1 - \eta} \;=\; 0. \tag{3.35}$$

□

Tout vecteur $x = (x_0, \ldots, x_{n-1})$ de $\mathbb{C}^n$ peut donc s'écrire

$$x \;=\; \sum_{j=0}^{n-1} c_j f_j, \tag{3.36}$$

où

$$c_j \;=\; (\, f_j \,|\, x \,) \;=\; \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} e^{-\frac{2i\pi}{n} \cdot jk} x_k \tag{3.37}$$

est un coefficient complexe. La donnée des n coefficients (c_i) caractérise entièrement le vecteur x et est appelée **transformée de Fourier discrète**.

Lien avec les familles précédemment définies Dans le cas où $n = 2$, on retrouve bien la famille orthonormale réelle définie équation (3.24). Mais pour les valeurs de n suivantes, les f_j sont complexes. On verra plus loin que ce n'est pas un problème : en un sens à préciser plus bas, le cadre complexe est bien le plus adapté à l'analyse théorique. Par contre, dans les algorithmes pratiques, il peut être utile d'avoir à disposition des bases d'harmoniques *réelles*. Après une étude théorique effectuée avec les harmoniques complexes, on privilégiera donc in fine des transformées réelles telles la base des cosinus discrets, projections sur $\mathbb{R}^n$ des vecteurs définis dans $\mathbb{C}^n$.

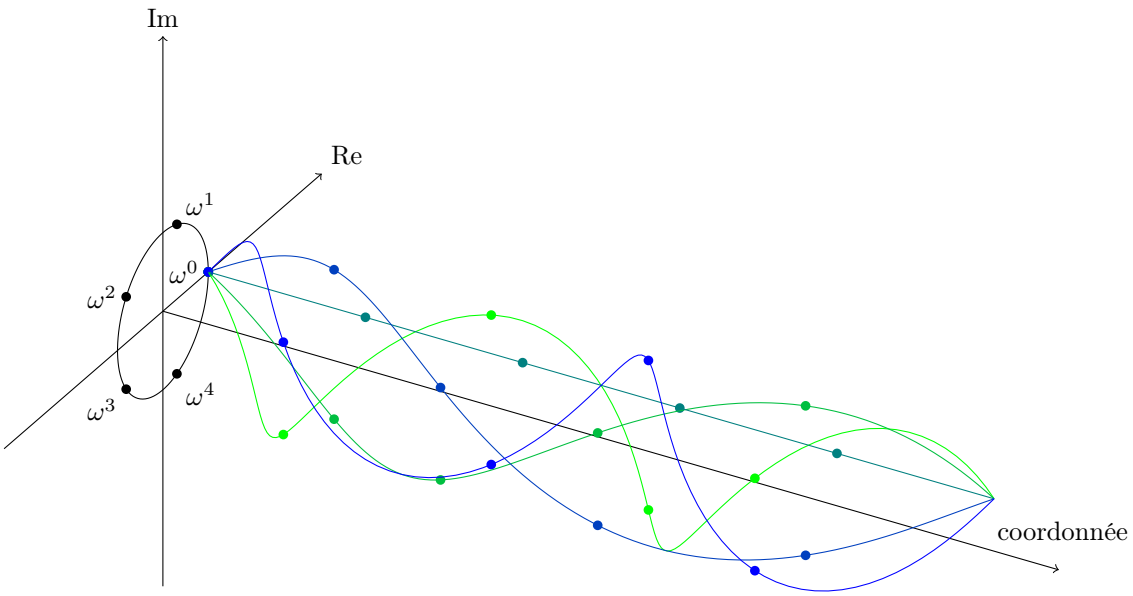


FIGURE 3.16 – Base $(f_0, f_1, f_2, f_3, f_4)$ des vecteurs harmoniques de longueur $n = 5$. À un facteur de normalisation $1/\sqrt{n}$ près, tous prennent leurs valeurs dans le "cercle discret" $\mathbb{U}_5 = \{1, \omega, \omega^2, \omega^3, \omega^4\}$, où $\omega = e^{\frac{2i\pi}{5}}$ est tel que $\omega^5 = 1$, mais le parcourent à des vitesses différentes : on a $f_a = (\omega^{a\cdot 0}, \omega^{a\cdot 1}, \omega^{a\cdot 2}, \omega^{a\cdot 3}, \omega^{a\cdot 4})$.
Cette figure met en valeur l'enroulement (discret) des harmoniques autour de l'origine du plan complexe, 0. Avec ses courbes d'interpolation entre coordonnées, elle préfigure le passage aux harmoniques *continues*.

Transformée de Fourier continue

Pourquoi nous limiter aux seuls espaces de dimension finie ? La Figure 3.16 doit nous inciter à passer des vecteurs de $\mathbb{C}^n$, (identifiables à des fonctions définies sur $[\![1, n]\!]$) à de véritables fonctions définies sur le continuum $\mathbb{R}$ tout entier.

Nous avons vu aux équations (3.2-3.3) que les valeurs $f(x)$ ne sont jamais que les coordonnées d'un vecteur "fonction" f dans la base canonique des diracs δ_x. De manière analogue à ce que nous avions fait équation (3.8), on peut donc définir le produit scalaire entre deux fonctions f et g définies de $\mathbb{R}$ à valeurs dans $\mathbb{C}$ par :

$$(f \,|\, g) \;=\; \int_{-\infty}^{+\infty} \overline{f(x)} \cdot g(x) \,\mathrm{d}x, \tag{3.38}$$

et on aura

$$\|f\|_2^2 \;=\; (f \,|\, f) \;=\; \int_{-\infty}^{+\infty} \overline{f(x)} \cdot f(x) \,\mathrm{d}x \;=\; \int_{-\infty}^{+\infty} |f(x)|^2 \,\mathrm{d}x. \tag{3.39}$$

Le passage à la dimension infinie impose de prendre quelques précautions techniques : on ne peut ainsi plus vraiment dire que la base des diracs est une "famille orthonormale" classique, puisque le produit de deux diracs n'est pas défini. Dans la suite du chapitre, je passerai néanmoins ces détails sous silence : il ne faudrait pas noyer les intuitions essentielles sous une flopée de mises en gardes peu parlantes.

La généralisation naturelle des harmoniques discrètes est la suivante :

Définition 3.4 (Transformée de Fourier)**.** Pour toute pulsation $\omega \in \mathbb{R}$, on définit l'harmonique associée

$$f_\omega : x \mapsto \frac{1}{\sqrt{2\pi}} e^{i\omega \cdot x}. \tag{3.40}$$

La famille des $(f_\omega)_{\omega \in \mathbb{R}}$ se comporte alors en tous points comme une base orthonormale :

— Toute fonction f raisonnable admet une *transformée de Fourier* $\widehat{f}$, donnée des produits scalaires entre f et les f_ω :

$$\forall\, \omega \in \mathbb{R},\; \widehat{f}(\omega) = (f_\omega \,|\, f) = \frac{1}{\sqrt{2\pi}} \int_{t=-\infty}^{+\infty} e^{-i\omega \cdot t} f(t) \,\mathrm{d}x. \tag{3.41}$$

— De manière analogue à ce qui est écrit équation (3.15), on peut reconstruire f à partir de $\widehat{f}$: pour presque tout $x \in \mathbb{R}$,

$$f(x) = \int_{\omega=-\infty}^{+\infty} \widehat{f}(\omega)\, f_\omega(x) \,\mathrm{d}\omega = \frac{1}{2\pi} \int_{\omega=-\infty}^{+\infty} \int_{t=-\infty}^{+\infty} f(t) e^{-i\omega \cdot t}\, e^{i\omega \cdot x} \,\mathrm{d}t \,\mathrm{d}\omega. \tag{3.42}$$

— L'identité de Parseval reste vérifiée. On a notamment

$$\int_{x=-\infty}^{+\infty} |f(x)|^2 \,\mathrm{d}x \;=\; \|f\|_2^2 \;=\; \int_{\omega=-\infty}^{+\infty} |\widehat{f}(\omega)|^2 \,\mathrm{d}x. \tag{3.43}$$

Normes 2 et produits scalaires se calculent aussi simplement avec les $\widehat{f}(\omega)$ que dans la base des diracs (δ_x). Dans les pages qui suivent, nous verrons que la base des harmoniques possède un immense avantage sur cette dernière : elle *diagonalise* la dérivation.

Interprétation fréquentielle Nous avions déjà rencontré les f_ω au chapitre 2, dans la preuve par homotopie du théorème fondamental de l'algèbre. Rappelez-vous : comme indiqué équation (2.36), $f_\omega : t \mapsto e^{i\omega t}$ n'est rien d'autre que le lacet qui parcourt le cercle unité à une **vitesse angulaire** ω, en partant de 0 au temps $t = 0$. Pour caractériser ce mouvement périodique, on pourra recourir selon les usages à trois termes équivalents : la **pulsation** ω ; la **période** $2\pi/\omega$, temps mis par $f_\omega(t)$ à effectuer un tour complet ; la **fréquence** $\omega/2\pi$, ou nombre de tours par unité de temps.

L'équation (3.42) exprime que toute fonction (raisonnable) d'une variable réelle peut être comprise comme une superposition de fonctions **harmoniques** élémentaires, les

$$\widehat{f}(\omega) f_\omega(\cdot) : x \mapsto \rho_\omega e^{i(\omega x + \theta_\omega)}, \tag{3.44}$$

où $\widehat{f}(\omega) = \rho_\omega e^{i\theta_\omega}$. Chacune de ces harmoniques, associée à une pulsation ω, est un lacet qui tourne en rond sur le cercle de rayon ρ_ω, à vitesse angulaire ω constante, après être parti d'un angle θ_ω à l'instant $t = 0$.

Dans cette décomposition, les $\widehat{f}(\omega) f_\omega(\cdot)$ associés à des pulsations ω positives tournent dans le sens trigonométrique, tandis que les pulsations ω négatives sont associées à des mouvements dans le sens des aiguilles d'une montre. On comprendra alors qu'un signal $f(x)$ est réel pour tout x si et seulement si on a, pour toute pulsation ω, $\widehat{f}(-\omega) = \mathrm{conj}(\widehat{f}(\omega))$ – exercice ! De même, plus une fonction est *régulière*, moins ses variations seront rapides. Elle n'aura donc pas besoin de faire entrer dans sa décomposition en harmoniques des fonctions f_ω oscillant trop rapidement : on peut montrer qu'une fonction est d'autant plus régulière que le module de sa transformée $|\widehat{f}(\omega)|$ décroît rapidement vers 0 lorsque $|\omega|$ tends vers $+\infty$.

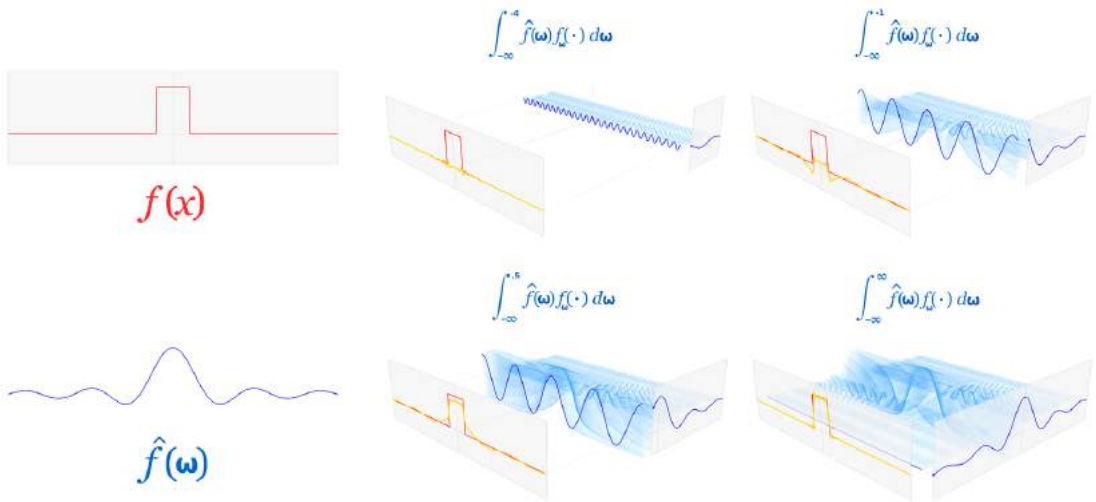


FIGURE 3.17 – Une fonction créneau comme somme de sa série de Fourier. Par commodité, on ne représente ici que des parties réelles : les harmoniques sont ici figurées par des fonctions sinusoïdales. Le signal, en rouge, est approché au fur et à mesure par les sommes partielles (en jaune) de sa représentation en harmoniques continues, équation (3.42).
Image adaptée de Wikipédia, par Lucas V. Barbosa.

Séries de Fourier Si f est une fonction périodique (disons, de période 2π pour simplifier), sa transformée $\widehat{f}$ prend une forme bien particulière. Impossible en effet d'y trouver des composantes de pulsations quelconques : seules les harmoniques f_ω qui sont elles-mêmes 2π-périodiques pourront entrer dans la décomposition. Or $f_\omega : x \mapsto e^{i\omega\cdot x}$ ne saura être 2π-périodique que si ω est entier, dans $\mathbb{Z}$. La transformée de Fourier de f prendra donc la forme d'un peigne

$$\widehat{f} \;=\; \sum_{\omega\in\mathbb{Z}} c_\omega \cdot \delta_\omega, \tag{3.45}$$

où les coefficients c_ω donnent la mesure de la corrélation entre f et f_ω sur une période :

$$c_\omega \;=\; \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} e^{-i\omega\cdot t} f(t)\,\mathrm{d}t. \tag{3.46}$$

Ces expressions localisées permettent de construire des coefficients finis, comparables entre eux, en rangeant l'infinité du nombre de périodes sur $\mathbb{R}$ dans le dirac δ_ω. Moralité : toute fonction périodique f est caractérisée par une suite de coefficients complexes c_ω et on peut écrire, pour presque tout t dans $\mathbb{R}$,

$$f(t) \;=\; \sum_{\omega\in\mathbb{Z}} c_\omega e^{i\omega\cdot t}. \tag{3.47}$$

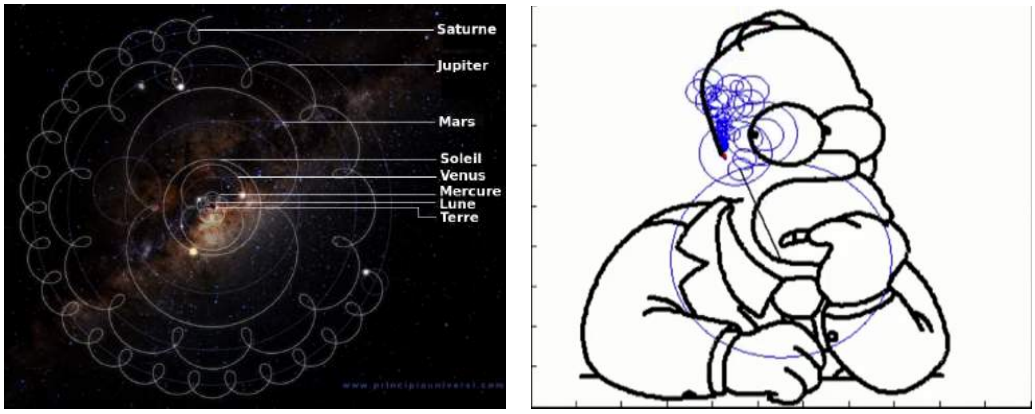
Ptolémée et ses épicycles : pas si bête ! Si l'on écrit $c_\omega = \rho_\omega e^{i\theta_\omega}$, cette identité prend un sens géométrique clair : tout chemin $f(t)$ à valeurs dans le plan complexe de période T est une somme de trajectoires circulaires, de fréquences multiples de la fréquence fondamentale $1/T$, caractérisées chacune par un rayon ρ_ω et un décalage angulaire θ_ω. C'est le principe qu'utilisèrent les astronomes de l'antiquité pour calculer les positions relatives des planètes : jusqu'au temps de Kepler, les orbites planétaires étaient décrites par des tables de coefficients "de Fourier". Déterminées empiriquement, elles permettaient de rendre compte des va-et-viens périodiques effectués par les planètes sur la voute céleste.

Si cette méthode n'a pas de limite théorique, elle peut vite se révéler encombrante : à mesure que la précision des mesures augmente, il faut ajouter de nouveaux coefficients qui rendent compte du fait que les trajectoires des corps célestes ne sont pas issues d'un grand *Spirograph* cosmique... Disposant de mesures d'une qualité exceptionnelle (effectuées par Tycho Brahe), Kepler proposera un modèle plus pratique qui repose sur des trajectoires *elliptiques*. Nous avons vu au chapitre 9 comment Newton avait pu en déduire les lois classiques de l'interaction gravitationnelle. Plus tard, c'est la détection d'une "anomalie" dans l'orbite de Mercure qui poussera Einstein à développer sa théorie de la relativité générale.

Pourquoi les nombres complexes ? La théorie des harmoniques est la deuxième grande motivation derrière l'étude des nombres complexes. Au lycée et dans les classes préparatoires, on insiste beaucoup sur son intérêt *algébrique* : les équations polynomiales, le théorème fondamental de l'algèbre auquel nous avons consacré la deuxième partie du chapitre 2. Mais l'ensemble des nombres complexes, c'est aussi le cadre privilégié pour l'étude des phénomènes *périodiques*. Impossible en effet de définir une fonction périodique de $\mathbb{R}$ à valeurs dans $\mathbb{R}$ qui fasse consensus. Créneaux, triangles, dents de scie... rien de tout cela n'est satisfaisant, ni même régulier.

A contrario, dans le plan complexe, un mouvement périodique s'impose par son évidence : celui qui consiste à **tourner en rond**. Il est précisément décrit par la famille des harmoniques $f_\omega : t \mapsto e^{i\omega\cdot t}$, que l'on fera redescendre dans $\mathbb{R}$ pour définir les *fonctions trigonométriques* :

$$\cos(\omega t) = \mathrm{Re}\big(e^{i\omega\cdot t}\big) = \frac{e^{+i\omega\cdot t} + e^{-i\omega\cdot t}}{2}, \qquad \sin(\omega t) = \mathrm{Im}\big(e^{i\omega\cdot t}\big) = \frac{e^{+i\omega\cdot t} - e^{-i\omega\cdot t}}{2i}. \tag{3.48}$$



(a) Modèle de Ptolémée pour le système solaire. (b) Portrait d'Homer Simpson, par Ptolémée.

FIGURE 3.18 – Tout signal périodique peut être décrit comme une somme de trajectoires circulaires. Les trajectoires dessinées en (a) vous rappelleront peut-être votre enfance : l'instrument de dessin *Spirograph* repose sur le même principe.
Images tirées des vidéos YouTube *Ptolemy's model of the universe* de Andrej Rehak et *Ptolemy and Homer (Simpson)* de Santiago Ginnobili, que vous pourrez trouver aux adresses `www.youtube.com/watch?v=EpSy0Lkm3zM` `www.youtube.com/watch?v=QVuU2YCwHjw`.

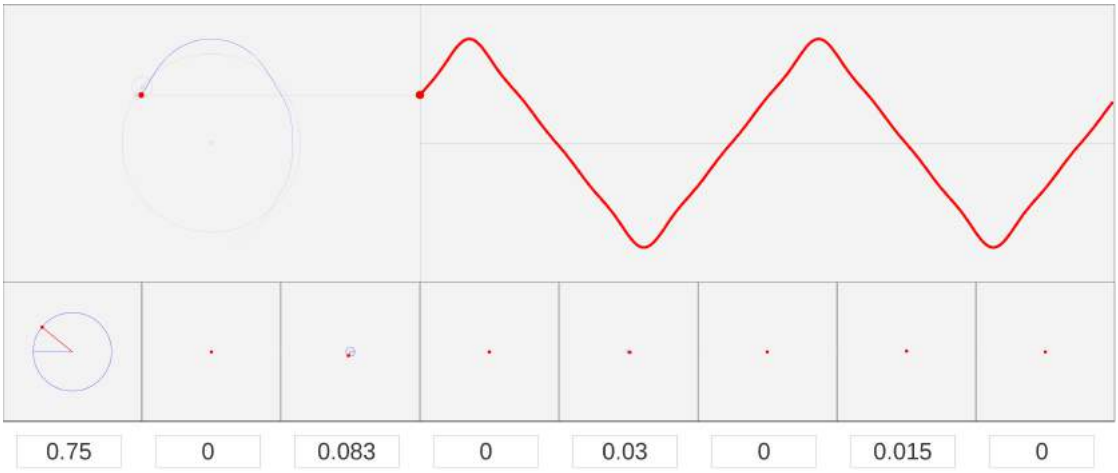


FIGURE 3.19 – Les séries de Fourier dans le plan complexe : Sur la ligne du bas, on peut sélectionner huit coefficients $c_1, \ldots, c_8$ représentés ici par les rayons de cercles. Chacun des points rouges parcourt ces cercles à une fréquence de 1 à 8 tours par seconde. La somme des huit vecteurs est alors affichée dans le panneau de gauche, l'origine faisant office de référentiel. L'ordonnée de ce point mobile complexe, signal réel périodique, est finalement tracée à droite comme sur un oscilloscope. Les coefficients ont ici été choisis pour coller au plus près d'un signal en triangle : assez régulier, il se laisse facilement approcher et nos huit coefficients suffisent à en obtenir une approximation très satisfaisante.
Travail de Lucas V. Barbosa, tiré du site `toxicdump.org/stuff/FourierToy.swf` que je vous encourage vivement à visiter – clic droit pour afficher les commandes et informations. Pour d'autres animations interactives sur la transformée de Fourier, on pourra consulter le site *BetterExplained*, `betterexplained.com/articles/an-interactive-guide-to-the-fourier-transform/`

Retour sur la compression JPEG On comprend maintenant d'où venait la famille des *cosinus discrets* utilisée par l'algorithme de compression JPEG – équation (3.6). À un décalage d'indice 1/2 près (introduit pour conserver une certaine parité après discrétisation), il s'agit précisément de la transposition aux signaux réels des harmoniques complexes de dimension 2, les

$$f_{\omega_1,\omega_2} : (x_1, x_2) \mapsto e^{i(\omega_1 x_1 + \omega_2 x_2)}. \tag{3.49}$$

Au prix de calculs roboratifs, on peut montrer que la base des cosinus discrets conserve de nombreuses propriétés de son parent continu (orthonormalité à un facteur près, analyse fréquentielle, etc.) tout en étant adaptée aux algorithmes numériques entiers. Il faut donc la comprendre comme une transposition fidèle d'un concept mathématique régulier (les harmoniques) au monde discret des ordinateurs.

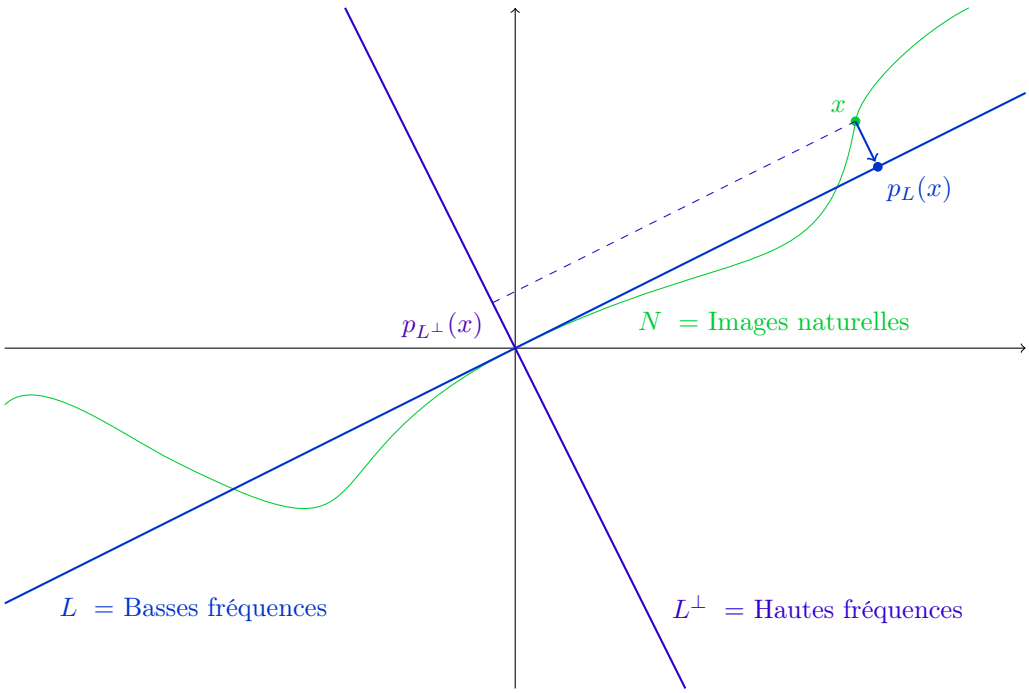


FIGURE 3.20 – Idée géométrique sous-jacente à la compression JPEG. Dans l'espace des images, l'ensemble N des *images naturelles* est tortueux, impossible à caractériser mathématiquement. On peut néanmoins tenter de l'approcher par un espace L plus simple : dans notre cas, celui des images *basses fréquences*. En première approximation, pour comprimer une image x quelconque, il suffit alors d'en considérer la *projection orthogonale* $p_L(x)$ sur L.
Si L est de faible dimension, on pourra décrire $p_L(x)$ par un faible nombre de coefficients, d'où la compression. Si L est suffisamment proche de N, la perte nette, ou résiduel $p_{L^\perp}(x) = x - p_L(x)$ sera négligeable. Trouver un espace L qui soit petit tout en approchant au mieux l'ensemble des images naturelles, c'est le défi posé mathématicien appliqué.
Cet effort de *modélisation* d'un ensemble d'images réelles est toujours poursuivi aujourd'hui – voir le format JPEG2000 illustré Figure 3.12, qui repose sur une théorie des *ondelettes* établissant un continuum entre les diracs δ_x et les ondes harmoniques f_ω. Cette théorie permet la compression, l'analyse ou le débruitage de signaux variés (données sismiques, images médicales, ...).

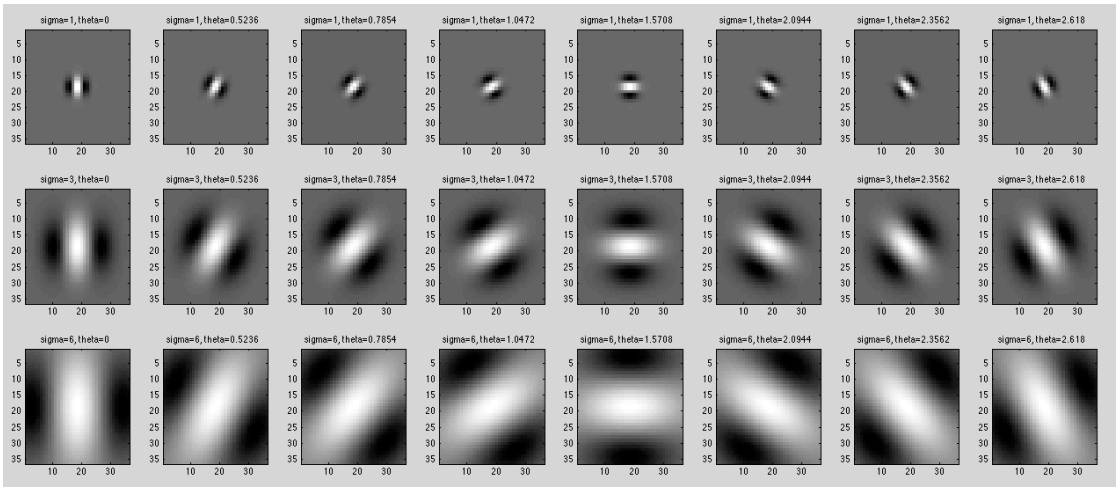


FIGURE 3.21 – Exemple de famille d'ondelettes orientées, en dimension 2. Localisées en position, fréquence (échelle) et orientation, ces images peuvent être comparées à des *notes* de musiques qui décrivent, de manière efficace, les images naturelles. On a représenté ici trois échelles (lignes) et 8 orientations (colonnes) ; en translatant chacune de ces images dans tout le domaine spatial, on obtient une famille suffisante pour une première analyse.
La recherche de bases orthonormales encore plus adaptées que celle des *cosinus discrets* au traitement d'image naturelles a été un sujet de recherche particulièrement dynamique au tournant des années 2000, avec le développement d'une zoologie foisonnante des familles d'ondelettes acceptables. Récompensée par le prix Abel décerné à Yves Meyer cette année, cette science peut être considérée comme "l'analyse harmonique du XXI^{e} siècle". Image tirée du site `eric-yuan.me/morlet-wavelet/`.

(a) JPEG2000. (b) Bandelettes. (c) Une photo d'identité sur code-barre.

FIGURE 3.22 – Un code-barre 2D typique peut contenir de l'ordre de 500 octets d'information. Grâce à des familles d'ondelettes orientées, les *bandelettes*, cela peut suffire à encoder de manière satisfaisante des visages, ou de petites images naturelles. En (a) et (b), on présente deux photos d'identité comprimées pour tenir sur un code barre de 500 octets (.5 Ko !) : de JPEG2000 (ondelettes séparables) à la décomposition en bandelettes, une nette amélioration a été apportée. C'est par exemple un moyen de sécuriser des badges d'identification, en stockant sous une forme cryptée facilement accessible une "copie de référence" du visage du propriétaire de la carte.
Fondée par Stéphane Mallat, la startup *Let it Wave* en a fait son fond de commerce et propose des puces électroniques effectuant l'opération de codage/décodage de manière efficace. Images tirées du site de M. Mallat, aujourd'hui professeur au Collège de France après être passé par le département d'informatique de l'École : `www.di.ens.fr/~mallat/papiers/CRM-Mallat-Course2.pdf`.

Une base adaptée à la dérivation

Depuis le début du chapitre, j'ai tenu à introduire les harmoniques f_ω de manière progressive. Pas à pas, nous avons pu établir qu'il s'agit d'une famille "orthonormale" de fonctions contenant la fonction constante égale à 1, f_0. Mais pourquoi cette construction, et pas une autre ? C'est que les harmoniques sont exactement les fonctions *bornées* pour lesquelles **la dérivation correspond à la multiplication par un coefficient scalaire fixé**.

Théorème 3.2 (Les harmoniques f_ω sont les vecteurs propres de la dérivation)**.** *Soit f une fonction bornée de $\mathbb{R}$ à valeurs dans $\mathbb{C}$. Alors les propositions suivantes sont équivalentes :*

1. *Il existe une constante λ dans $\mathbb{C}$ telle que $f' = \lambda f$.*
2. *Il existe une constante μ dans $\mathbb{C}$ et une pulsation ω dans $\mathbb{R}$ telle que $f = \mu f_\omega$.*

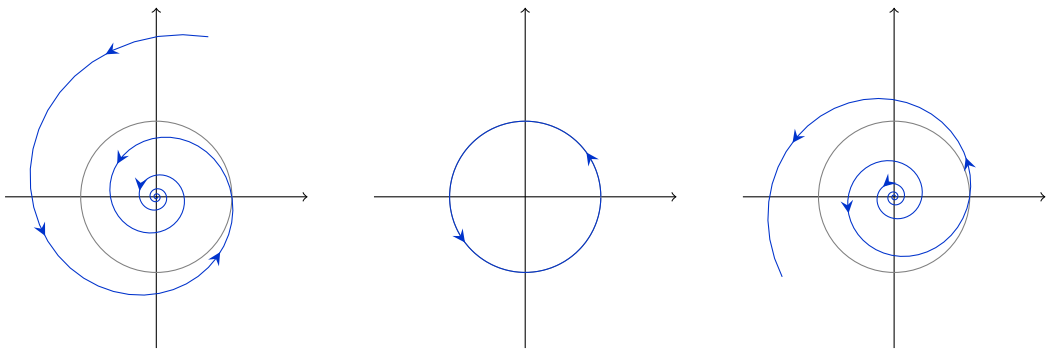
Démonstration. Que *2* implique *1* est clair : si $f = \mu f_\omega$, on a

$$\forall\, x \in \mathbb{R}, f'(x) = \mu \frac{\mathrm{d}}{\mathrm{d}x}(e^{i\omega x}) = i\omega\mu e^{i\omega x} = i\omega f(x). \tag{3.50}$$

Réciproquement, supposons qu'il existe $\lambda \in \mathbb{C}$ tel que $f' = \lambda f$. D'après le théorème de Cauchy-Lipschitz sur les équations différentielles, on sait alors qu'il existe μ dans $\mathbb{C}$ tel que

$$\forall\, x \in \mathbb{R}, f(x) = \mu e^{\lambda x}. \tag{3.51}$$

Mais alors, pour que f soit bornée sur $\mathbb{R}$ (conformément à l'hypothèse de départ), il est indispensable que la partie réelle de λ soit nulle. C'est donc que l'on peut écrire $\lambda = i\omega$, où $\omega \in \mathbb{R}$. □



(a) $\mathrm{Re}(\lambda) < 0$. (b) $\mathrm{Re}(\lambda) = 0$. (c) $\mathrm{Re}(\lambda) > 0$.

FIGURE 3.23 – Si $f(t)$ est un point courant du plan complexe, l'équation $f'(t) = \lambda f(t)$ définit exactement une spirale logarithmique : la tangente à la courbe au point M fait un angle constant $\theta = \arg(\lambda)$ avec le segment $[OM]$, où O est l'origine du repère. La courbe est parcourue à une vitesse angulaire constante égale à $\mathrm{Im}(\lambda)$.
Pour qu'une telle courbe reste bornée dans le plan, il est nécessaire que λ soit d'argument $\pm\pi/2$, i.e. que λ soit un imaginaire pur.

Dans la foulée, on peut démontrer que la transformée de Fourier rend *diagonal* l'opérateur de dérivation. Là où calculer la dérivée d'une fonction donnée dans la "base des diracs" par ses valeurs ponctuelles demandait un effort certain (passage à la limite des taux d'accroissements, etc.), la même opération devient triviale dans la base des harmoniques de Fourier.

Théorème 3.3 (Dérivation et transformée de Fourier, énoncé informel)**.** *Soit f une fonction définie sur $\mathbb{R}$ à valeurs dans $\mathbb{C}$ – par exemple de carré intégrable – dont on peut considérer la transformée de Fourier $\widehat{f}$. Alors sous réserve d'existence, on a*

$$\forall\ \omega \in \mathbb{R},\ \widehat{f'}(\omega) \ = \ i\omega \widehat{f}(\omega). \tag{3.52}$$

Dans la base d'harmoniques, la dérivation peut donc se comprendre comme une ***simple multiplication par un facteur $i\omega$, qui dépend de la composante****.*

Démonstration. Par définition de la transformée de Fourier, on peut écrire f comme une combinaison linéaire continue des harmoniques $f_\omega : x \mapsto e^{i\omega x}$:

$$\forall\ x \in\ \mathbb{R},\ f(x) \ = \ \int_{\omega=-\infty}^{+\infty} \widehat{f}(\omega) \cdot e^{i\omega x}\, \mathrm{d}\omega. \tag{3.53}$$

Mais alors, sous réserve d'hypothèses techniques peu restrictives sur f, on peut dériver cette somme terme à terme :

$$\begin{aligned} \forall\ x \in\ \mathbb{R},\ f'(x) \ &= \ \int_{\omega=-\infty}^{+\infty} \widehat{f}(\omega) \cdot i\omega e^{i\omega x}\, \mathrm{d}\omega & (3.54)\\ &= \ \int_{\omega=-\infty}^{+\infty} \widehat{f'}(\omega) \cdot e^{i\omega x}\, \mathrm{d}\omega & (3.55) \end{aligned}$$

Par unicité de l'écriture de f' dans la base des f_ω, on obtient donc le résultat. □

Si le théorème précédent demande un peu de doigté technique (les sommations continues ne sont pas aussi simples à manipuler que les sommes discrètes), son énoncé doit rester intuitif : la dérivation agit sur une fonction f comme un filtre *passe-haut*, qui tue les basses fréquences ($|\omega|$ petit) et rehausse les oscillations de faible amplitude ($|\omega|$ grand).

Utilisation pratique Dans $\mathbb{C}^n$, on a l'habitude des opérations *diagonales* qui agissent séparément sur chaque coordonnée : ce sont simplement les *changements d'échelles*. L'application

$$S : x = (x_{-2}, x_{-1}, x_0, x_1, x_2) \mapsto (-2 \cdot x_{-2}, -1 \cdot x_{-1}, 0, 1 \cdot x_1, 2 \cdot x_2) \tag{3.56}$$

n'effarouche maintenant plus personne. Eh bien, dériver une fonction f, ce n'est pas plus compliqué que cela. Il suffit de passer du système de coordonnées $(f(x))_{x\in\mathbb{R}}$ à la transformée $(\widehat{f}(\omega))_{\omega\in\mathbb{R}}$, que l'on peut identifier à un vecteur de $\mathbb{R}^{\mathbb{R}}$. On multiplie alors chacune de ses coordonnées par un facteur $i\omega$, comme on avait multiplié x_n par un facteur n dans l'équation ci-dessus. Reste à repasser d'une représentation fréquentielle à un graphe $(f'(x))_{x\in\mathbb{R}}$, et le tour est joué !

Décomposition dans une base d'harmoniques

Ce théorème en poche, nous pouvons enfin aborder le problème dont l'étude poussa Joseph Fourier (1768-1830) à développer une théorie moderne des harmoniques : la résolution analytique d'équations aux dérivées partielles.

Équation de la chaleur Dans sa *Théorie analytique de la chaleur* parue en 1822, Fourier propose de modéliser l'évolution d'un champ de température sur un fil de fer par l'équation dite "de la chaleur" :

$$\forall\, x \in \mathbb{R}, \forall\, t \geqslant 0, \quad \frac{\partial T}{\partial t}(x,t) \;=\; D\frac{\partial^2 T}{\partial x^2}(x,t), \tag{3.57}$$

où $T(x,t)$ est la température au point d'abscisse x du fil à l'instant T, et D un coefficient de diffusion thermique. Nous avions déjà introduit cette équation au chapitre précédent, équation (10.62), et discuté de son sens physique. Grâce aux travaux de Galerkine et à la méthode des éléments finis, on peut la résoudre numériquement sur tout domaine de l'espace (le volume d'une tasse en céramique, par exemple). Mais comment comprendre des solutions purement calculatoires ? De manière complémentaire à la seule résolution de problèmes d'ingénierie, le scientifique veut *donner du sens* au monde qui l'entoure. Pouvoir décrire, en quelques phrases, un phénomène physique.

Réécriture dans le domaine fréquentiel Dans notre cas, voir cette évolution dans la base des harmoniques f_ω va considérablement simplifier l'interprétation du phénomène. Supposons disposer d'un fil de fer infini, et notons

$$T_t : x \in \mathbb{R} \mapsto T(x,t) \in \mathbb{R} \tag{3.58}$$

le champ de température à l'instant t. La *condition initiale*, imposée à l'instant $t = 0$, est donc simplement la donnée de T_0. On considère alors la transformée $\widehat{T}_t$, et on écrit la répartition de chaleur comme une superposition d'harmoniques de toutes amplitudes :

$$T(x,t) = T_t(x) = \int_{\omega=-\infty}^{+\infty} \widehat{T}_t(\omega)\, f_\omega(x)\, \mathrm{d}x. \tag{3.59}$$

Mais alors, l'équation de la chaleur devient simplement :

$$\forall\, x \in \mathbb{R}, \forall\, t \geqslant 0, \quad \int_{\omega=-\infty}^{+\infty} \frac{\partial\Big(\widehat{T}_t(\omega)\Big)}{\partial t}\, f_\omega(x)\, \mathrm{d}x = \int_{\omega=-\infty}^{+\infty} (i\omega)^2\, D\, \widehat{T}_t(\omega)\, f_\omega(x)\, \mathrm{d}x \tag{3.60}$$

puisque dériver deux fois par rapport à x, c'est multiplier la transformée de Fourier par $(i\omega)\cdot(i\omega)$. Par unicité de la décomposition dans la base d'harmoniques, l'équation (3.57) s'écrit donc, dans le domaine de Fourier :

$$\forall\, \omega \in \mathbb{R}, \forall\, t \geqslant 0, \quad \frac{\partial\Big(\widehat{T}_t(\omega)\Big)}{\partial t} = -\omega^2\, D\, \widehat{T}_t(\omega). \tag{3.61}$$

La méthode de Fourier L'évolution du *vecteur* de dimension infinie T_t au cours du temps est maintenant comprise. Dans les coordonnées "de Dirac" $(T_t(x))_{x\in\mathbb{R}}$, l'équation de la chaleur était intuitive (un minimum local de température se réchauffe, un maximum se refroidit), mais difficile à résoudre. Par contre, elle est *diagonale* dans les coordonnées harmoniques $(\widehat{T}_t(\omega))_{\omega\in\mathbb{R}}$: chaque coordonnée reste indépendante des autres, évolue en roue libre selon l'équation différentielle à un paramètre (3.61), "$\dot{y} = -\omega^2 D y$". Regarder les équations d'évolution de fonctions dans la base des harmoniques, c'est la remarquable leçon donnée par Joseph Fourier.

Résolution et interprétation physique Dans ce cas précis, les solutions à l'équation d'évolution des coefficients de Fourier sont connues : on trouve simplement que

$$\forall\, \omega \in \mathbb{R}, \forall\, t \geqslant 0, \quad \widehat{T_t}(\omega) = \widehat{T_0}(\omega)\, e^{-\omega^2\, D\cdot t}, \tag{3.62}$$

ce qui caractérise l'évolution du champ de température au cours du temps.

On sait que $t \mapsto e^{-\omega^2\, D\cdot t}$ est une fonction qui part de 1 pour tendre vers 0 à l'infini, avec un temps caractéristique de décroissance en $1/D\omega^2$. L'interprétation physique du phénomène décrit par l'équation de la chaleur est donc la suivante :

« Les harmoniques de T_0 s'atténuent de façon exponentielle, d'autant plus vite qu'elles correspondent à des fréquences élevées. »

Partant d'une répartition de chaleur arbitraire T_0, on va donc observer une *régularisation* du profil T_t, jusqu'à une *dissipation* complète de la chaleur initiale dans le fil de fer. Si la température moyenne $\widehat{T_t}(0)$ est conservée, le profil de la distribution de chaleur n'en finit pas moins par s'aplatir en une bosse qui s'étale en tendant vers 0.

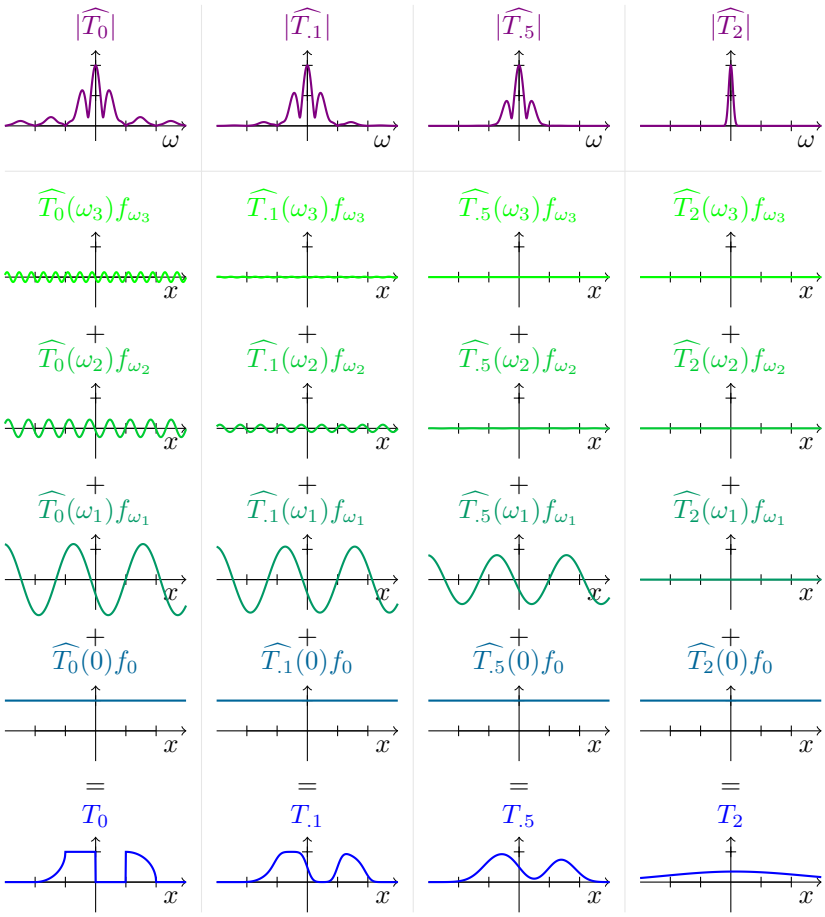


FIGURE 3.24 – Évolution du "vecteur de dimension infinie" T_t suivant l'équation de la chaleur. De haut en bas, on voit : la transformée de Fourier de T_t à l'instant t, représentée par son module ; 4 harmoniques prépondérantes de T_t, qui correspondent aux pics de $|\widehat{T_t}|$ dans le cadran droit : $\omega = 0$ (la bosse du milieu), $\omega = \omega_1$ pour la bosse de seconde amplitude, puis $\omega = \omega_2$ et ω_3 pour les répliques hautes fréquences qui permettent d'encoder les discontinuités de T_0 ; enfin, sur la dernière ligne, le graphe de T_t, profil de la température sur le fil de fer.
De gauche à droite, on peut observer l'évolution de ces représentations au cours du temps, à quatre instants $t = 0$, .1, .5 et 2. Le simple affaissement coordonnée à coordonnée de $\widehat{T_t}(\omega)$ décrit équation (3.62), particulièrement rapide dans les hautes fréquences, permet donc bien de retrouver le phénomène de dissipation de la chaleur.

Équation des cordes vibrantes De même, on peut poursuivre l'étude de l'équation (10.58) de D'Alembert, qui décrit le comportement d'une corde de violon de longueur L :

$$\frac{\partial^2 y}{\partial t^2} = v^2 \frac{\partial^2 y}{\partial x^2}, \tag{3.63}$$

où $y(t,x) = y_t(x)$ est l'altitude de la corde au temps t à l'abscisse x, et où $v = \sqrt{T/\mu}$ est une vitesse caractéristique dépendant de la masse linéique μ de la corde et de la tension T exercée à ses extrémités.

Analyse de Fourier Ici, on a imposé la condition :

$$\text{pour tout instant } t, \quad y_t(0) \;=\; 0 \;=\; y_t(L). \tag{3.64}$$

Les seules harmoniques pouvant intervenir dans la décomposition de y_t sont donc celles qui sont associées à des longueurs d'onde en rapport entier avec $2L$, c'est à dire aux pulsations spatiales ω multiples de la **pulsation fondamentale** $\omega_{\mathbf{L}} = \pi/\mathbf{L}$. In fine, de manière analogue à la décomposition (3.47) sur les fonctions périodiques, on peut donc écrire

$$\forall\, x \in [0,L],\; y_t(x) \;=\; \sum_{n\in\mathbb{Z}} c_t(n) f_{n\cdot\omega_L}(x). \tag{3.65}$$

Mais alors, par linéarité de l'équation de D'Alembert et en utilisant le fait que $\frac{\partial}{\partial x} f_\omega = i\omega f_\omega$, on trouve que le profil y_t de la courbe vérifie l'équation des cordes vibrantes si et seulement si

$$\forall\, n \in \mathbb{Z}, \quad \frac{\partial^2 (c_t(n))}{\partial t^2} = -n^2\omega_L^2 v^2 c_t(n). \tag{3.66}$$

Il s'agit de l'équation des cordes, ré-exprimée dans les coefficients de Fourier.

Résolution Dans le plan complexe, chaque coefficient $c_t(n)$ dépendant du temps vérifie une équation du type $\ddot{z} = -\lambda z$ que l'on peut résoudre simplement. En supposant par exemple que la corde est lâchée sans vitesse à l'instant $t = 0$, on a :

$$\forall\, n \in \mathbb{Z}, \forall\, t \geqslant 0, \quad c_t(n) \;=\; c_0(n) \cdot \cos(n\omega_L v \cdot t). \tag{3.67}$$

En décomposant en harmoniques *spatiales* le profil initial de la corde, on a trouvé des coefficients $c_0(n)$ qui correspondent à des profils sinusoïdaux à n bosses entre $x = 0$ et $x = L$. L'équation ci-dessus nous garantit simplement que ces différentes composantes "superposées" dans le profil de la courbe au temps t **ne vont pas interagir entre elles** : elles se contentent d'*osciller* sagement, avec une pulsation temporelle $n\omega_L v = n\pi v/L$ et donc une fréquence temporelle $|n|v/2L$ proportionnelle au nombre de bosses de l'harmonique considérée. En pratique, ce mouvement d'oscillation haute fréquence (20Hz $\sim$ 20 000Hz) est accompagné d'une atténuation progressive liée aux frottements et à la dissipation de l'énergie dans l'air (temps de décroissance de l'ordre de la seconde, bien supérieur à la période des oscillations).

Interprétation physique Conséquence : la vibration de la corde au cours du temps est une superposition de vibrations harmoniques élémentaires de fréquences temporelles $|n|v/2L$, multiples de la *fréquence fondamentale* $v/2L$. En admettant une certaine linéarité de la caisse de résonance de l'instrument (typiquement valable pour des vibrations de faible amplitude), on a donc retrouvé une propriété générale des instruments de musique, illustrée Figure 3.26 :

« En régime libre, une corde ou un tube de longueur L produiront une vibration d'air que l'on peut écrire comme une superposition de signaux périodiques – les *harmoniques* du son – associés à des fréquences multiples de la *fréquence fondamentale* du système, celle-ci inversement proportionnelle à L toutes propriétés égales par ailleurs (tension de la corde, diamètre du tube...). »

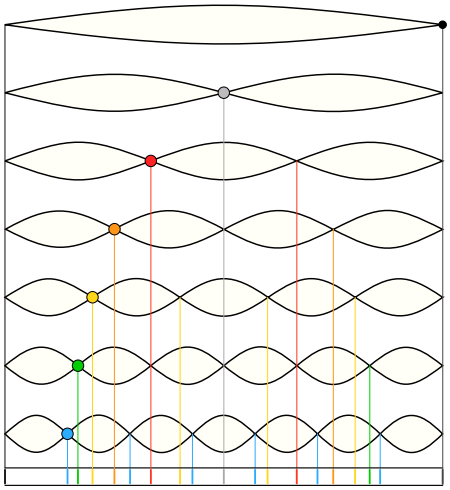


FIGURE 3.25 – Premières harmoniques spatiales d'une corde de violon de longueur L. L'équation (3.67) assure que celles-ci vibrent indépendamment les unes des autres, à une fréquence inversement proportionnelle à leurs longueurs d'ondes spatiales respectives : $\mathbf{f} = \mathbf{nv}/\mathbf{2L}$, où n est le nombre de bosses qui va ici de 1 à 7.
Image tirée de Wikipédia, par Moodswingerscale.

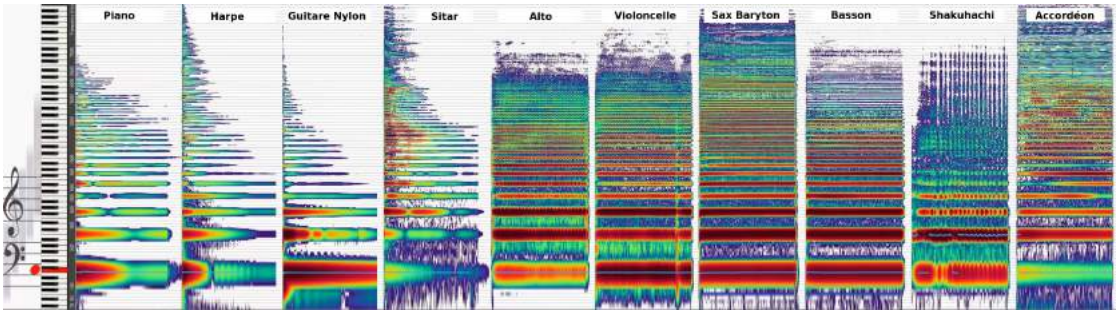


FIGURE 3.26 – Spectrogrammes pour divers instruments de musique jouant un *do* grave. En abscisse, on représente le temps : les notes sont tenues pendant cinq bonnes secondes, et se succèdent les unes après les autres. En ordonnée, on représente le module de la transformée de Fourier *instantanée* au temps t : avec l'axe des pulsations ω orienté de bas en haut selon l'échelle *logarithmique* de la partition, cette figure permet d'estimer le contenu fréquentiel d'un son assimilé à une fonction périodique sur chaque intervalle $[t, t + \Delta t]$, où Δ_t est une résolution temporelle. Plus la couleur est chaude, plus l'énergie $|\widehat{f}_t(\omega)|^2$ portée par le son au temps t à la fréquence ω est importante.
Première leçon : quel que soit l'instrument, la note produite est portée par les même fréquences multiples de la fréquence fondamentale correspondant au *do* grave, avec une pulsation $\omega_0 = 2\pi{\cdot}130.813$ Hz. En bonne approximation, ces fréquences correspondent aux notes do grave (ω_0), do medium ($2\,\omega_0$), sol medium ($3\,\omega_0$), do aigu ($4\,\omega_0$), mi aigu ($5\,\omega_0$), sol aigu ($6\,\omega_0$), etc. Si les bandes paraissent écrasées, c'est parce que l'axe des ordonnées est en graduation logarithmique, à l'instar des oreilles qui perçoivent les *rapports* et non les *différences* de fréquences.
À l'oreille, seuls les *modules* des coefficients de Fourier instantanés importent. Le timbre de l'instrument est donc déterminé entièrement par le spectrogramme ci-dessus, qui dépend de deux grands paramètres : la répartition de l'énergie sur l'échelle des multiples de ω_0 ; la tenue de la note, ou la dissipation de l'énergie après l'excitation initiale du système (pincement de la corde, impact du marteau...).
En tous cas, l'analyse physique de la page précédente se trouve confortée expérimentalement.
Image adaptée de la vidéo *Timbre : why different instruments playing the same tone sound different* de la chaîne *What Music Really Is* : `www.youtube.com/watch?v=VRAXK4QKJ1Q`.

Principe de fonctionnement d'une flûte Régie par une équation du même type (la pression de l'air dans le tube y jouant le rôle de la hauteur de la corde), la flûte permet d'illustrer parfaitement ces comportements, la décomposition du son en harmonique superposées. Il s'agit essentiellement d'un tube dont les propriétés physiques (matériau, diamètre) déterminent le *timbre*, tandis que la *note* jouée est conséquence des actions du musicien aux bords de la colonne d'air.

D'un côté, une *embouchure* en biseau où le souffle du musicien crée un tourbillon limite. De l'autre, des trous qui permettent aux doigts de l'instrumentiste de faire varier la *longueur* L de la colonne d'air prisonnière du tube. On peut alors expliquer la production de notes de la manière suivante :

> « En faisant varier L, le musicien choisit la fréquence fondamentale de résonance du tube $f_L = k/L$, où k est un paramètre fixe dépendant de la perce et du bois. Excitée par un tourbillon limite à l'entrée du tube, la colonne d'air va adopter un comportement périodique de fréquence f_L, va-et-vient de la surpression le long du tube qui va alternativement faire entrer ou repousser le tourbillon d'air de manière cyclique. En sortie du tube, on obtiendra une surpression périodique de l'air de fréquence fondamentale f_L : la note de musique. »

À la flûte baroque (sans clés), c'est ainsi que l'on produit les notes dites du *premier registre* : observez sur la gauche de la Figure 3.32 comme les notes ré, mi$^\flat$, fa$^\flat$, fa$^\sharp$, sol, la et si sont simplement obtenues en levant un à un les doigts des deux mains. Les notes intercalaires sont quand à elles produites par des doigtés "de fourche", qui tirent parti du faible diamètre des trous : même ouverts, ceux-ci ne libèrent pas complètement la colonne d'air. En ouvrant un trou tout en fermant les suivants, on réussit à obtenir un comportement "entre-deux", identifiable à une longueur de tube intermédiaire.

Le son produit est périodique, de fréquence fondamentale $f_L = k/L$: notre oreille est faite de telle sorte que l'on identifiera cette fréquence "la plus basse" comme la *note* jouée, *enrichie* par les harmoniques supérieures associées à des fréquences multiples de f_L. C'est ce que l'on observe sur les spectrogrammes de la Figure 3.26, où toutes les notes jouées sont perçues comme des do graves aux timbres variables.

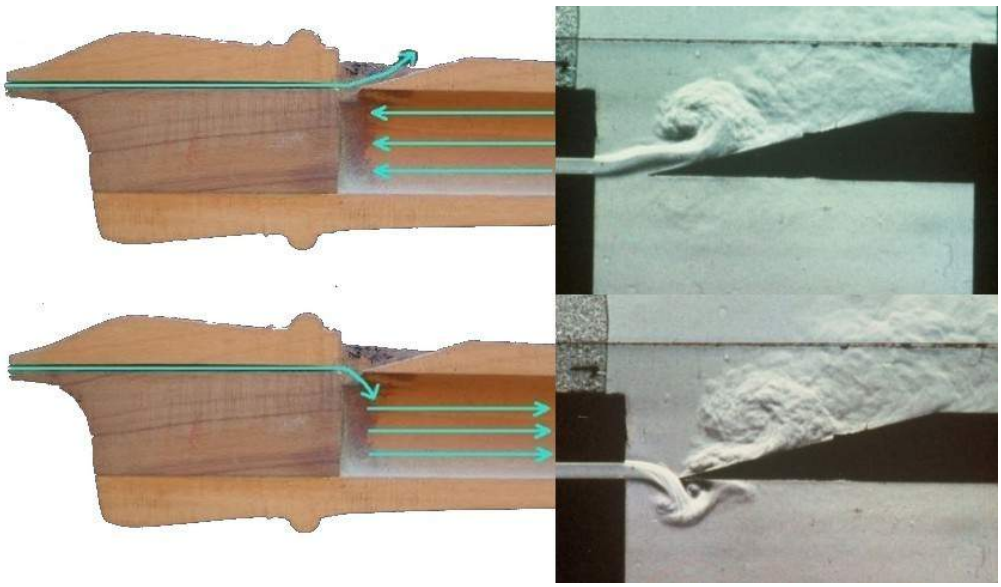


FIGURE 3.27 – Illustration du cycle tourbillonnant à l'embouchure d'une flûte à bec. Le cycle sortant/entrant se répète à une fréquence déterminée par les propriétés acoustiques du tube. Figure tirée du site de Philippe Bolton, facteur de flûtes `www.flute-a-bec.com/acoustique.html` – les images de droite sont issues de l'article *Luchtwervels in een blokfluit*, Avraham Hirschberg, université d'Eindhoven.

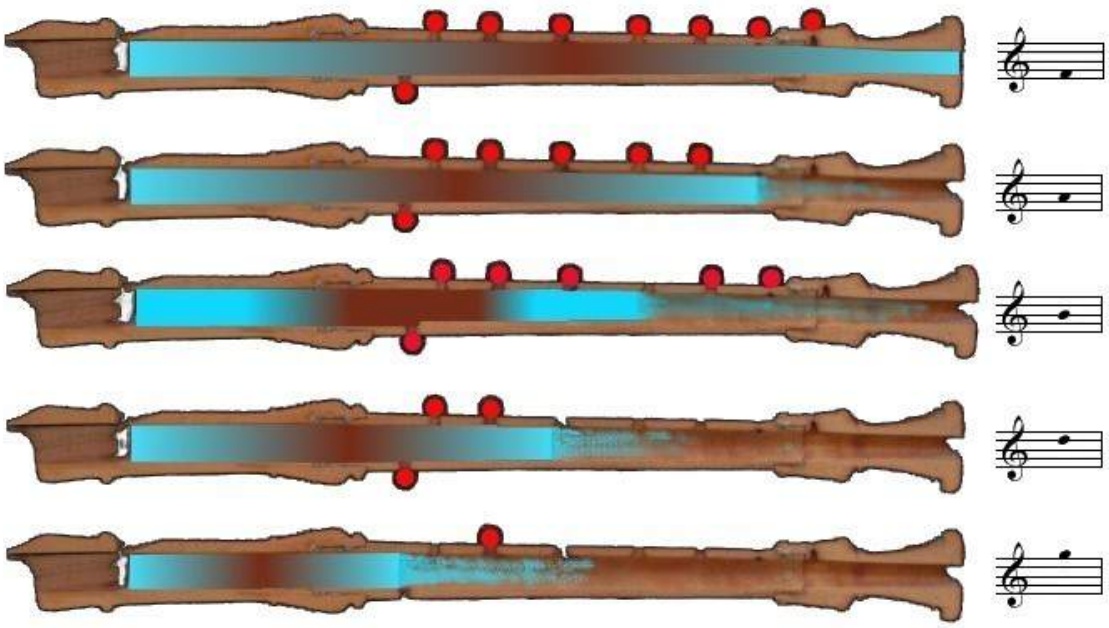


FIGURE 3.28 – Nœuds de pression (harmonique fondamentale) correspondant à différents doigtés sur une flûte à bec alto baroque en *fa* jouant dans le **premier registre** : le bleu correspond à une dépression, le rouge à une surpression et les disques rouges marquent les trous fermés. De haut en bas, on trouve le fa, la et si graves, ré et sol médiums, notes qui sonnent d'autant plus aigu que la colonne d'air en vibration est courte.
Figure tirée du site de Philippe Bolton.

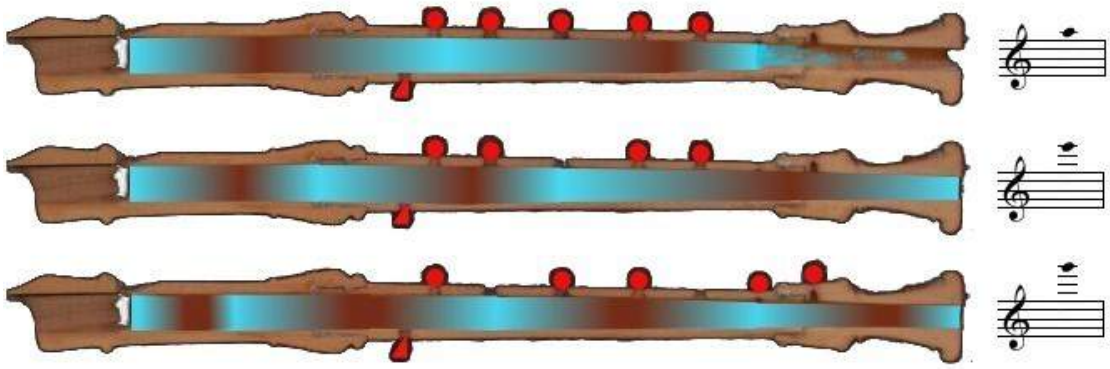


FIGURE 3.29 – Illustration du **deuxième, troisième et quatrième registres** d'une flûte à bec en fa. En ouvrant partiellement le trou du pouce et en ouvrant certains trous au milieu du corps de l'instrument, on impose des conditions du type "$y_t(x_i) = 0$" pour tout temps en certaines positions de trous x_i, ce qui élimine les harmoniques basses fréquences de l'écriture du champs de pression variable $y_t(x)$. Ceci permet d'enlever au son sa fréquence fondamentale, pour n'en retenir que les harmoniques de plus petites longueurs d'ondes, associées à des notes aiguës : ici, un la médium, mi aigu et sol aigu.
Figure tirée du site de Philippe Bolton.

Extraction des harmoniques Jouer dans le *premier registre* d'une flûte, c'est se contenter de faire sonner l'harmonique fondamentale du tube. Le principe physique dégagé par l'équation (3.67), valable pour tout les instruments, est que **la fréquence fondamentale de vibration de l'air est inversement proportionnelle à la longueur L du tube**. Le couper en deux, c'est donc multiplier par deux la fréquence, monter d'une octave ; le couper en quatre, c'est, pareillement, monter de deux octaves.

Alors, comment réaliser en pratique un instrument à la tessiture étendue qui puisse jouer sur deux, trois gammes à la fois ? Sur un piano, il suffit d'aligner les cordes côte à côte, en filetant les plus graves pour les alourdir et s'éviter ainsi d'avoir un rapport de 1 à 2^{10} entre les longueurs des cordes de droite (aiguës, donc courtes) et celles de gauche (associées aux notes graves, plus longues).

Mais à la flûte, quel embarras ! Impossible de jouer les plus belles pièces du répertoire s'il faut changer d'instrument à chaque fois que l'on change de gamme... Heureusement, pour jouer des notes aiguës sur un tube *médium* standard, il existe deux astuces.

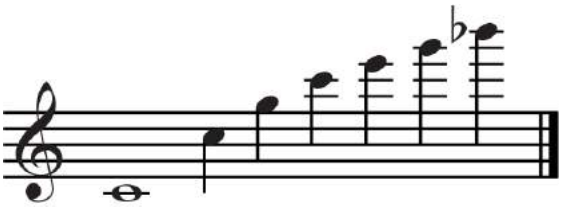
Doigtés D'abord, comme à la flûte à bec, on peut choisir d'ouvrir certains trous pour imposer des conditions d'annulation "$y_t(x_i) = 0$" à la surpression de l'air dans le tube, où x_i est l'abscisse du trou laissé délibérément ouvert. Par exemple, entre la deuxième ligne de la Figure 3.28 et la première ligne de la Figure 3.29, on a laissé entrouvert le trou du pouce gauche, ce qui impose une dépression au milieu de la perce. En interdisant la surpression "marron" associée au la grave, on divise par deux la longueur d'onde spatiale du champs de pression dans le tube, et on double donc la fréquence "fondamentale" de vibration de l'instrument. Sans surprise, on est passé du la grave au la médium.

Vitesse du souffle À la flûte traversière, le musicien peut agir sur un autre levier : la vitesse de son souffle. Sans changer de doigté, il lui est possible de *couper les harmoniques*, en interdisant au tube de vibrer selon les fréquences les plus basses. Si, par exemple, l'harmonique fondamentale du tube f_0 était associée à un do grave (au milieu du clavier de piano ; correspond à un tube complètement fermé sur une flûte traversière moderne en do), souffler de plus en plus vite va permettre d'égrener une à une les harmoniques du son de fréquences f_0, $2\,f_0$, $3\,f_0$, $4\,f_0$, $5\,f_0$, $6\,f_0$ en tronquant le spectrogramme par le bas, ce qui s'entendra comme une suite do grave, do médium, sol médium, do aigu, mi aigu, sol aigu de plus en plus appauvrie. Il s'agit d'un exercice de son classique, illustré Figure 3.31.

Le système Boehm Pour jouer une mélodie ambitieuse utilisant à fond les harmoniques "non fondamentales" de son instrument, le flûtiste semble donc devoir choisir entre des doigtés complexes (difficiles à enchaîner) et un contrôle subtil et permanent de son souffle (ce qui complique grandement l'expression musicale). C'est sans compter sur le travail de Theobald Boehm, flûtiste bavarois qui mis au point un système de clefs très élaboré, entre 1831 et 1847. En rendant accessible les doigtés de fourche les plus extravagants, celui-ci limite l'utilisation de la vitesse du souffle aux seuls changements d'octaves, en simplifie considérablement la montée-descente chromatique des notes de la gamme. Il s'agit aujourd'hui du système de référence, monté sur toutes les flûtes et clarinettes modernes : on comprend maintenant pourquoi.

Bilan L'analyse ci-dessus ne prétends pas être parfaitement rigoureuse : en extrapolant à partir de l'équation de D'Alembert (valable au premier ordre pour les cordes de violons), nous avons mis sous le tapis toutes les questions relatives à la pression et aux conditions limites imposées par les orifices du tube. Néanmoins, cette petite étude nous aura permis de comprendre qualitativement l'acoustique des instruments à vent. Elle aura eu le mérite de démystifier les tablatures que l'on fait d'ordinaire apprendre aux enfants. Armé d'une méthode efficace de résolution des équations différentielles, du langage des harmoniques de Fourier, le physicien est donc capable d'expliquer simplement un des phénomènes les plus intrigants de notre vie quotidienne : beau succès !

FIGURE 3.30 – Quelques instruments à vent pour illustrer cette fin de chapitre. Tout en haut, on trouve un *bawu* chinois, sorte de flûte à anche libre très facile à jouer mais d'une tessiture restreinte, sans possibilité de grimper dans les harmoniques. Ensuite, deux *dizi* (flûtes traversières) en bambou : une petite, aiguë et une grande jouant en médium. Notez sur chacune d'elles la fine membrane, entre l'embouchure et les trous, qui change les "conditions aux bords" de la colonne d'air et confère à ces flûtes une sonorité bien particulière. En 4e et 5e position, un *traverso* baroque vendu avec deux corps : le premier, plus long, qui permet de jouer au diapason La=415Hz (comme au temps de Bach) et le second, plus court, accordé aux instruments modernes avec un La=440Hz. Enfin, tout en bas, une *flûte traversière* moderne équipée d'un système Boehm.



(a) Harmoniques obtenues à partir du do.

(b) Exercice à écouter !

FIGURE 3.31 – En maintenant le doigté du do grave (tous les trous fermés) et en soufflant de plus en plus vite dans l'embouchure, on obtient les premières harmoniques du do. Je vous encourage à écouter le résultat dans la vidéo *Six Flute Harmonics on Middle C* de David Fei, `www.youtube.com/watch?v=TmkKQ_DpXkI`, et à le comparer aux fichiers sons analogues sur wikipédia, `en.wikipedia.org/wiki/File:Violin_harmonics.ogg` (harmoniques en La du violon) et `en.wikipedia.org/wiki/File:Harmonics_110x16.ogg` (harmoniques sinusoïdales sur le la grave, aux spectrogrammes purs).

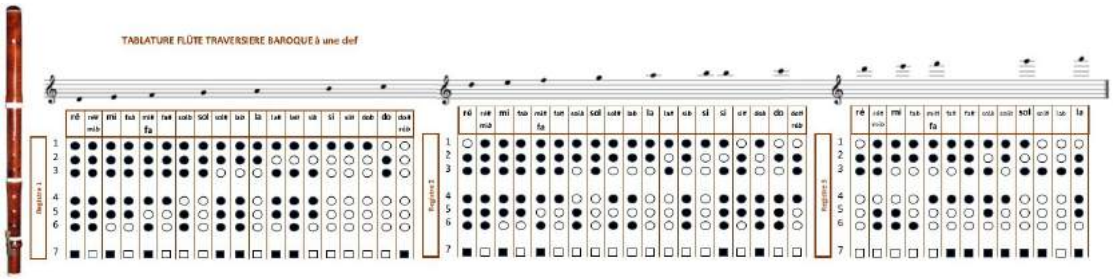


FIGURE 3.32 – Tablature pour un traverso baroque en *do*. Un point noir correspond à un trou fermé, un point blanc à un doigt levé. La clé du bas est un usage de commodité, qui permet à l'auriculaire droit d'atteindre le dernier trou.
Image tirée du site `www.flutesbaroques.com/tablature.php`.

Conclusion

Récapitulons. Dans ce chapitre plus que dans tout autre, nous avons pu apprécier le regard *utile* et *original* que les mathématiques portent sur le monde. Parler d'espaces de "dimension infinie", ce n'est pas s'abandonner à une délirante fantaisie métaphysique. C'est, simplement, admettre la complexité de certains problèmes posés par le continu ; et sans se décourager, tenter de les résoudre en se reposant sur un vocabulaire connu, celui des droites et des plans.

Accepter de dire que les images, les cordes ou les orbites planétaires sont des *vecteurs*, données d'une infinité de coordonnées, c'est accepter de dépasser les nombres pour faire de la *géométrie*. Alors certes, une image en pratique, ce n'est pas une "infinité" de coordonnées : juste un très gros paquet, de 256^2 nombres par exemple. Mais comment pourrait-on se satisfaire d'une représentation "pixel-à-pixel" qui s'étend à mesure que la résolution des capteurs s'améliore ? Attaquer de front le problème "limite" de la dimensions infinie, organiser l'information de manière sémantique (des basses et hautes fréquences qui font sens, pas une mosaïque de diracs asservies aux cellules de nos appareils), c'est voir percer la structure derrière les suites de coordonnées.

Fait étonnant, le même langage des harmoniques oscillantes nous est apparu dans des domaines a priori étrangers les uns aux autres :

— La mécanique céleste des anciens, avec le système de Ptolémée.

— La théorie de la musique, des violons et des flûtes.

— La théorie de la chaleur et de sa diffusion.

— La transmission moderne d'images, le format JPEG.

Et ce sans même parler de l'électronique ou des lois de l'optique. Aujourd'hui, ce point de vue harmonique s'est généralisé à toute l'analyse, a même diffusé en algèbre. Si nous nous sommes attardé en cours sur les harmoniques de la droite réelle, les f_ω, physiciens et mathématiciens n'ont de cesse aujourd'hui de décrire leurs équivalents dans des espaces moins homogènes (une tasse en céramique, un cristal de roche...) ou de dimension supérieure.

Le premier à avoir systématisé ce travail, "compris d'une manière exacte et complète la nature des séries trigonométriques" (pour citer Riemann), Joseph Fourier aura donc initié un mouvement qui touche autant aux théories mathématiques qu'aux applications pratiques. Homme ancré dans son temps, au service de ses contemporains (il fut préfet de l'Isère de 1802 à 1815), il est un modèle pour de nombreux mathématiciens appliqués. Mais la reconnaissance fut longue à venir. À ce sujet, on rapporte souvent la controverse déclenchée par Jacobi écrivant en 1830 à Legendre :

> « M. Fourier avait l'opinion que le but principal des mathématiques était l'utilité publique et l'explication des phénomènes naturels ; mais un philosophe comme lui aurait dû savoir que le but unique de la science, c'est l'*honneur de l'esprit humain*, et que sous ce titre, une question de nombres vaut autant qu'une question du système du monde. »

Reprise par Dieudonné, Bourbaki, la formule fait mouche ; on y lit entre les lignes la condescendance pour les applications qui imprègne toujours l'esprit de nombreux étudiants en mathématiques "pures". Jusqu'en 1970, pas un article dans l'Encyclopédie Universalis au nom de Joseph Fourier ! Heureusement, les temps changent ; nous assistons aujourd'hui à une remise à l'honneur des nombreux mathématiciens qui ont voulu, en plus de faire progresser la science, se rendre utiles à leurs prochains – voir *Le retour de Fourier* par Jean-Pierre Kahane, `www.academie-sciences.fr/pdf/dossiers/Fourier/Fourier_pdf/Fourier_Kahane.pdf`.

Tous comptes faits, si je ne devais conserver qu'une leçon de ce cours, ce serait celle-ci :

« Pour aller de l'avant, changez de repère ! »

Chapitre 4

Introduction à la géométrie Riemannienne

Séance 5

Lors de la dernière séance, nous avons envisagé la géométrie à la manière d'un élève en licence de mathématiques. Dépassant les a priori d'Euclide et consorts, nous n'avons pas hésité à parler de droite, d'orthogonalité dans des espaces de trajectoires, de signaux et d'images. En appliquant le vocabulaire et les concepts de la géométrie à des espaces de grande dimension, nous avons pu apprivoiser ces derniers : les intuitions sur les plans, les projections ont donné naissance à des algorithmes tout à fait non-triviaux, et ô combien utiles.

Mais, si nous avons eu l'audace de plaquer le bon vieux vocabulaire euclidien sur des partitions et des photos, nous n'avons jamais *remis en question* les concepts géométriques sous-jacents à la géométrie des anciens : droite engendrée copie de la droite réelle et structure additive globale.

Et après tout, qui pourrait nous en blâmer ? Le monde de Descartes, l'espace euclidien $\mathbb{R}^3$ semble si naturel qu'on a du mal à l'imaginer autrement : entre deux points $a = (a_1, a_2, a_3)$ et $b = (b_1, b_2, b_3)$ passe un unique plus court chemin, le segment

$$s_{a\to b} : t \in [0,1] \mapsto (1-t)\cdot a + t \cdot b, \tag{4.1}$$

de longueur

$$\ell(a \to b) = \|a - b\| = \sqrt{(a_1-b_1)^2 + (a_2-b_2)^2 + (a_3-b_3)^2} \tag{4.2}$$

et voilà tout, il n'y a pas à chercher plus loin...

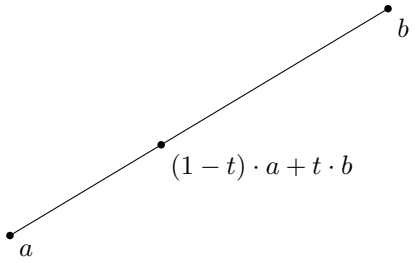


FIGURE 4.1 – Segment entre deux points de l'espace euclidien $\mathbb{R}^3$.

Axiomatiques non-euclidiennes

Vraiment ? Eh bien, plutôt qu'un long discours, attaquons ensemble l'entame du chapitre III de *La Science et l'Hypothèse*, d'Henri Poincaré.

> Les géométries non euclidiennes — Toute conclusion suppose des prémisses ; ces prémisses elles-mêmes ou bien sont évidentes par elles-mêmes et n'ont pas besoin de démonstration, ou bien ne peuvent être établies qu'en s'appuyant sur d'autres propositions, et comme on ne saurait remonter ainsi à l'infini, toute science déductive, et en particulier la géométrie, doit reposer sur un certain nombre d'axiomes indémontrables. Tous les traités de géométrie débutent donc par l'énoncé de ces axiomes. Mais il y a entre eux une distinction à faire : quelques-uns, comme celui-ci par exemple : « deux quantités égales à une même troisième sont égales entre elles », ne sont pas des propositions de géométrie, mais des propositions d'analyse. Je les regarde comme des jugements analytiques a priori, je ne m'en occuperai pas.
>
> Mais je dois insister sur d'autres axiomes qui sont spéciaux à la géométrie. La plupart des traités en énoncent trois explicitement :
>
> 1° Par deux points ne peut passer qu'une droite ;
>
> 2° La ligne droite est le plus court chemin d'un point à un autre.
>
> 3° Par un point on ne peut faire passer qu'une parallèle à une droite donnée.
>
> Bien que l'on se dispense généralement de démontrer le second de ces axiomes, il serait possible de le déduire des deux autres et de ceux, beaucoup plus nombreux, que l'on admet implicitement sans les énoncer, ainsi que je l'expliquerai plus loin.
>
> On a longtemps cherché en vain à démontrer également le troisième axiome, connu sous le nom de postulatum d'Euclide. Ce qu'on a dépensé d'efforts dans cet espoir chimérique est vraiment inimaginable. Enfin au commencement du siècle et à peu près en même temps, deux savants, un Russe et un Hongrois, Lobatchevsky et Bolyai établirent d'une façon irréfutable que cette démonstration est impossible ; ils nous ont à peu près débarrassés des inventeurs de géométries sans postulatum ; depuis lors l'Académie des Sciences ne reçoit plus guère qu'une ou deux démonstrations nouvelles par an.
>
> La question n'était pas épuisée ; elle ne tarda pas à faire un grand pas par la publication du célèbre mémoire de Riemann intitulé : Ueber die Hypothesen welche der Geometrie zum Grunde liegen. Cet opuscule a inspiré la plupart des travaux récents dont je parlerai plus loin et parmi lesquels il convient de citer ceux de Beltrami et de Helmholtz.
>
> La Géométrie de Lobatchevsky. — S'il était possible de déduire le postulatum d'Euclide des autres axiomes, il arriverait évidemment qu'en niant le postulatum, et en admettant les autres axiomes, on serait conduit à des conséquences contradictoires ; il serait donc impossible d'appuyer sur de telles prémisses une géométrie cohérente.
>
> Or c'est précisément ce qu'a fait Lobatchevsky. Il suppose au début que :
>
> L'on peut par un point mener plusieurs parallèles à une droite donnée ;

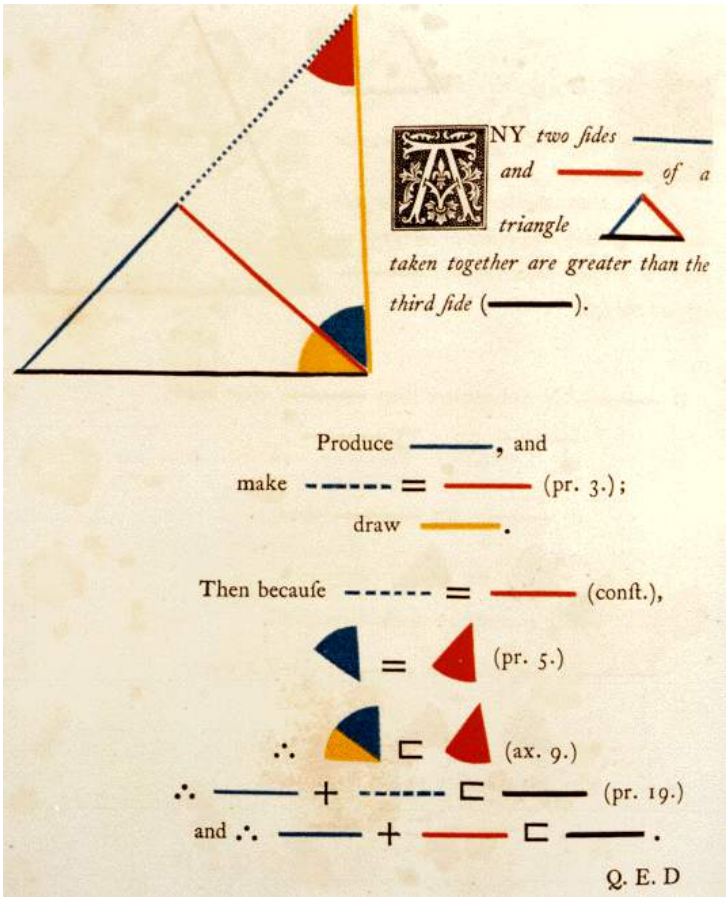


FIGURE 4.2 – Preuve par Euclide que la ligne droite est le plus court chemin entre deux points du plan (Livre 1, Proposition 20) – Illustration tirée de l'excellent *The Elements of Euclid*, d'Oliver Byrne, paru en 1848 et réédité aux éditions Taschen.

Et il conserve d'ailleurs tous les autres axiomes d'Euclide. De ces hypothèses, il déduit une suite de théorèmes entre lesquels il est impossible de relever aucune contradiction et il construit une géométrie dont l'impeccable logique ne le cède en rien à celle de la géométrie euclidienne.

Les théorèmes sont, bien entendu, très différents de ceux auxquels nous sommes accoutumés et ils ne laissent pas de déconcerter un peu d'abord.

Ainsi la somme des angles d'un triangle est toujours plus petite que deux droits et la différence entre cette somme et deux droits est proportionnelle à la surface du triangle.

Il est impossible de construire une figure semblable à une figure donnée mais de dimensions différentes.

Si l'on divise une circonférence en n parties égales, et qu'on mène des tangentes aux points de division, ces n tangentes formeront un polygone si le rayon de la circonférence est assez petit ; mais si ce rayon est assez grand, elles ne se rencontreront pas.

Il est inutile de multiplier ces exemples ; les propositions de Lobatchevsky n'ont plus aucun rapport avec celles d'Euclide, mais elles ne sont pas moins logiquement reliées les unes aux autres.

Une géométrie non-euclidienne

On l'a vu : d'autres *géométries*, d'autres jeux d'axiomes que celui retenu par Euclide sont donc cohérents. Mais comment nous représenter ces mondes – on préférera dire : ces espaces – aux règles inattendues ? Il suffit de pousser la lecture de *La Science et l'Hypothèse* jusqu'au chapitre suivant :

> Le monde non euclidien. — Si l'espace géométrique était un cadre imposé à chacune de nos représentations, considérée individuellement, il serait impossible de se représenter une image dépouillée de ce cadre, et nous ne pourrions rien changer à notre géométrie.
>
> Mais il n'en est pas ainsi, la géométrie n'est que le résumé des lois suivant lesquelles se succèdent ces images. Rien n'empêche alors d'imaginer une série de représentations, de tout point semblables à nos représentations ordinaires, mais se succédant d'après des lois différentes de celles auxquelles nous sommes accoutumés.
>
> On conçoit alors que des êtres dont l'éducation se ferait dans un milieu où ces lois seraient ainsi bouleversées pourraient avoir une géométrie très différente de la nôtre.

Henri Poincaré soulève ici un point crucial : personne n'a jamais *vu* de système de coordonnées attaché à l'espace ambiant. Il n'y a pas de *repère* intrinsèque nécessaire, de réalité des axes que l'on dessine par commodité dans les ouvrages de physique.

Celui-ci n'est qu'une construction a posteriori, qui reflète un fait cinématique bien établi dans notre vie quotidienne :

— Si j'avance de trois pas en avant, que je tourne à droite d'un tiers de tour (120°)...

— Puis que je ré-avance de trois pas, et que je re-tourne à droite d'un tiers de tour...

— Et qu'enfin, pour la troisième fois, j'avance de trois pas...

Eh bien, je serai revenu à mon point de départ ! Cette "vérité" de notre monde, c'est exactement celle qui est modélisée par la proposition suivante :

« Un triangle est équilatéral si et seulement si deux de ses angles sont d'un sixième de tour. »

Ni plus, ni moins. Une géométrie n'est pas une propriété consubstantielle à la matière, mais un résumé des règles auxquelles obéissent le mouvement. Sans altérer les propriétés des objets présents sur les "photos", il est donc possible d'imaginer de nouvelles géométries en modifiant le déroulement des vues sur la pellicule, les règles de transition d'un état à un autre.

FIGURE 4.3 – À gauche, illustration de la petite promenade décrite ci-dessus dans le plan euclidien. On pourrait tout à fait imaginer qu'une telle suite d'instructions ne nous ramène pas à notre point de départ, comme "illustré" sur la figure de droite : ce sera le point de départ des géométries non-euclidiennes.

Géométrie euclidienne, droites et géodésiques On s'en doute, la notion centrale sera ici celle de *droite*, au sens de plus court chemin reliant deux points. Avant de généraliser la notion d'espace géométrique, il est donc important de comprendre comment définir et calculer la longueur d'une courbe au sens de la géométrie euclidienne.

Étant donnés deux points a et b du plan euclidien, on s'intéressera aux *chemins* γ allant de a vers b, c'est-à-dire aux applications lisses du segment $[0,1]$ vers le plan telles que

$$\gamma(0) = a \qquad \text{et} \qquad \gamma(1) = b. \tag{4.3}$$

Le plus simple des chemins reliant $a = (a_1, a_2)$ à $b = (b_1, b_2)$ n'est autre que le segment

$$s_{a\to b} : t \mapsto (1-t)\cdot a + t \cdot b = \Big((1-t)\cdot a_1 + t\cdot b_1, (1-t)\cdot a_2 + t\cdot b_2\Big), \tag{4.4}$$

paramétré à vitesse constante, et dont la longueur est donnée par le théorème de Pythagore :

$$\ell_{\text{eucl}}(s_{a\to b}) = \sqrt{(b_1-a_1)^2 + (b_2-a_2)^2}. \tag{4.5}$$

En toute généralité, la longueur euclidienne d'un chemin lisse quelconque sera alors simplement définie comme la somme des longueurs des accroissements infinitésimaux $\mathrm{d}\gamma = \dot\gamma\,\mathrm{d}t$:

$$\ell_{\text{eucl}}(\gamma) = \int_0^1 \|\dot\gamma(t)\|\,\mathrm{d}t \tag{4.6}$$

$$= \int_0^1 \sqrt{\dot\gamma_1^2(t) + \dot\gamma_2^2(t)}\,\mathrm{d}t. \tag{4.7}$$

Fait remarquable : on peut entièrement caractériser les transformations du plan qui préservent la longueur des chemins – les *isométries* du plan euclidien.

Théorème 4.1 (Isométries du plan euclidien)**.** *Soit $f : \mathbb{R}^2 \to \mathbb{R}^2$ une application du plan euclidien dans lui-même. Alors les deux propositions suivantes sont équivalentes :*

1. *f préserve la longueur des chemins, au sens où, pour tout chemin $\gamma : [0,1] \to \mathbb{R}^2$,*
$$\ell(f\circ\gamma) = \ell(\gamma). \tag{4.8}$$
2. *f est la composée d'une translation, d'une rotation et, éventuellement, d'une réflexion, i.e. il existe un centre $c = (c_1, c_2)$, un vecteur $w = (w_1, w_2)$ et un angle θ tels que*
$$\forall\ (x_1,x_2)\in\mathbb{R}^2, f(x_1,x_2) = \begin{pmatrix} +(x_1-c_1)\,\cos(\theta) \pm (x_2-c_2)\,\sin(\theta) + c_1 + w_1 \\ -(x_1-c_1)\,\sin(\theta) \pm (x_2-c_2)\,\cos(\theta) + c_2 + w_2 \end{pmatrix} \tag{4.9}$$
ou, en notation complexe ($z = x_1 + i\,x_2$) :
$$\forall\ z\in\mathbb{C}, f(z) = \begin{cases} (z-c)\,e^{i\theta} + c + w & \textit{sans réflexion,} \\ \overline{(z-c)}\,e^{i\theta} + c + w & \textit{avec réflexion.} \end{cases} \tag{4.10}$$

Démonstration. L'implication réciproque est immédiate. Pour le sens direct, par contre, c'est un peu plus compliqué : on se ramène à devoir montrer qu'une isométrie f qui fixe le point $(0,0)$ (en quotientant les deux degrés de liberté accordé par la translation) et le point $(1,0)$ (en quotientant le degré de liberté accordé par la rotation) est ou bien l'identité, ou bien la réflexion d'axe (Ox).

Pour tout point $x = (x_1, x_2)$ du plan, on sait alors que

$$\|f(x) - (0,0)\| = \|x - (0,0)\| \qquad \|f(x) - (1,0)\| = \|x - (1,0)\|. \tag{4.11}$$

Autrement dit, $f(x)$ est à l'intersection des deux cercles non-concentriques qui contraignent exactement $f(x)$ à être égal ou bien à x, ou bien à son symétrique par rapport à l'axe horizontal. La détermination de l'image d'un point en dehors de l'axe entraînant celle des autres, on a bien démontré notre résultat : à une translation et rotation près, f est ou bien l'identité, ou bien la réflexion d'axe (Ox). □

Géodésiques du plan euclidien

Pour assurer la cohérence de notre terminologie, reste à démontrer un résultat des plus fondamentaux : pour joindre un point a à un point b, la droite (i.e. le segment $s_{a\to b}$) est le plus court chemin. Il s'agit donc de prouver le théorème suivant :

Théorème 4.2 (Caractérisation des géodésiques du plan euclidien)**.** *Soit $a = (a_1, a_2)$ et $b = (b_1, b_2)$ deux points du plan euclidien. Alors pour tout chemin γ allant de a à b, on a*

$$\ell_{eucl}(\gamma) \geqslant \ell_{eucl}(s_{a\to b}) = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2}, \tag{4.12}$$

avec cas d'égalité si et seulement si γ est égal au segment $s_{a\to b}$ à reparamétrisation près.

Démonstration. Pour démontrer ce théorème, on procède en deux temps. D'abord, on prouve le lemme suivant :

Lemme 4.1 (Principe de rétraction dans le plan euclidien)**.** *Si γ est un chemin joignant le point $(0,0)$ au point $(x,0)$, alors*

$$\ell(\gamma) \geqslant \ell_{eucl}(s_{(0,0)\to(x,0)}) = |x|\,, \tag{4.13}$$

avec cas d'égalité si et seulement si γ est équivalent à $s_{(0,0)\to(x,0)}$ à reparamétrisation près.

Lemme de rétraction. On écrit $\gamma(t) = (\gamma_1(t), \gamma_2(t))$. Quitte à reparamétrer γ, on peut supposer que $\|\dot\gamma(t)\| = \sqrt{\dot\gamma_1^2(t) + \dot\gamma_2^2(t)}$ est constant au cours du temps. C'est un point un peu technique à vérifier – nous ne le ferons pas ici : la paramétrisation par longueur d'arc. Autrement dit, on peut toujours parcourir le support d'une courbe à vitesse constante, ce qui ne change rien au calcul de la longueur. On a alors

$$\ell^2(\gamma) = \left(\int_0^1 \sqrt{\dot\gamma_1^2(t) + \dot\gamma_2^2(t)}\,\mathrm{d}t\right)^2 \tag{4.14}$$

$$= \int_0^1 \dot\gamma_1^2(t) + \dot\gamma_2^2(t)\,\mathrm{d}t \tag{4.15}$$

$$= \int_0^1 \dot\gamma_1^2(t)\,\mathrm{d}t + \int_0^1 \dot\gamma_2^2(t)\,\mathrm{d}t. \tag{4.16}$$

On notera que la paramétrisation par longueur d'arc est essentielle pour le passage de la première à la deuxième ligne, car il assure que la quantité intégrée est constante au cours du temps. Cette égalité doit être comprise comme un théorème de Pythagore sur les chemins, la décomposition du calcul de la longueur sur les deux axes (Ox) et (Oy) du repère.

Le point clé est alors de comprendre qu'entre ces deux composantes, la première est *utile* (puisqu'elle permet de passer de $\gamma_1(0) = 0$ à $\gamma_1(1) = x$), tandis que la seconde ne l'est pas (puisque par hypothèse $\gamma_2(0) = \gamma_2(1) = 0$).

Par suite, un chemin optimal restera nécessairement collé à l'axe horizontal avec $\gamma_2(t) = 0$ pour tout instant t. Conclure que le chemin optimal $\gamma = (\gamma_1, 0)$ est nécessairement égal au segment $s_{(0,0)\to(x,0)}$ n'est alors pas difficile (puisque l'on n'a jamais intérêt à rebrousser chemin entre 0 et x), et on a démontré notre lemme. □

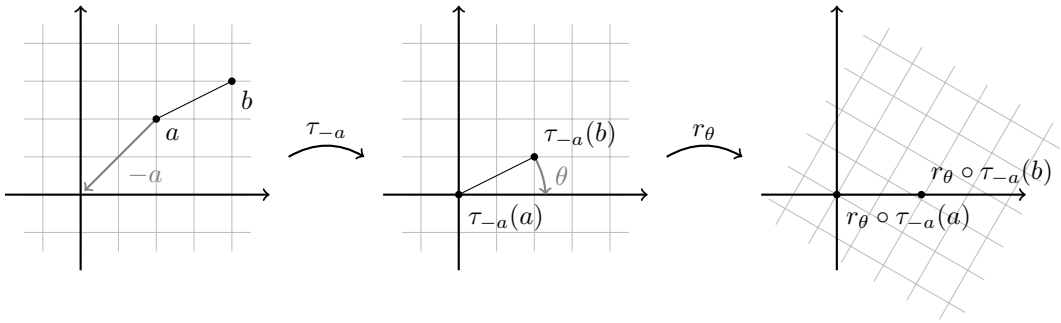


FIGURE 4.4 – Transport via une isométrie $f = r_\theta \circ \tau_{-a}$ d'une situation quelconque vers le cas "horizontal" traité par le lemme de rétraction.

Suite de la preuve Pour démontrer le théorème 4.2, une idée forte est alors d'utiliser les *isométries* du plan euclidien pour nous ramener au cas simple "$a = (0,0)$, $b = (x,0)$" traité explicitement par le lemme 4.1.

Considérons donc un chemin γ joignant le point a au point b, que l'on cherche à comparer au segment $s_{a\to b}$. D'après le théorème 4.1, on dispose d'une isométrie f telle que

$$f(a) = (0,0), \qquad f(b) = (x,0), \qquad \ell_{\text{eucl}}(f \circ \gamma) = \ell_{\text{eucl}}(\gamma) \tag{4.17}$$

avec $x = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2}$: il suffit de translater par $(-a)$, puis de considérer la bonne rotation. Fait remarquable, on a alors :

$$f \circ s_{a\to b} = s_{(0,0)\to(x,0)}. \tag{4.18}$$

On peut donc appliquer notre lemme au chemin $f \circ \gamma$ de manière pertinente :

$$\ell_{\text{eucl}}(\gamma) \;=\; \ell_{\text{eucl}}(f \circ \gamma) \;\geqslant\; \ell_{\text{eucl}}(s_{(0,0)\to(x,0)}) \;=\; \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2}, \tag{4.19}$$

avec cas d'égalité si et seulement si

$$f \circ \gamma = s_{(0,0)\to(x,0)} = f \circ s_{a\to b} \tag{4.20}$$

i.e. $\gamma = s_{a\to b}$ à reparamétrisation près. Cqfd. □

Bilan Tout autant que le résultat, il importe de retenir la démarche qui nous a permis de l'obtenir. Plutôt que de développer, calculer comme des forcenés la longueur d'un chemin arbitraire, nous avons découpé le problème en :

— Un calcul simple, dans un cas de référence (le lemme de rétraction).

— Un problème de "recalage", le transport d'une situation quelconque vers le cas de référence au moyen d'un *isomorphisme*, une transformation de l'espace qui conserve les propriétés étudiées – ici, les longueurs de chemins, d'où le nom plus spécifique d'*isométrie*.

L'exemple du monde sphérique

La méthode "géométrique" peut s'appliquer sans peine aux espaces homogènes "non plats", comme la sphère. Considérons par exemple, dans l'espace euclidien $\mathbb{R}^3$, la sphère unité

$$S^2 = \left\{(x_1, x_2, x_3) \in \mathbb{R}^3 \ , \ x_1^2 + x_2^2 + x_3^2 = 1\right\}, \tag{4.21}$$

dont nous isolons spécifiquement l'équateur

$$S^1 = \left\{(x_1, x_2, 0) \in \mathbb{R}^3 \ , \ x_1^2 + x_2^2 = 1\right\}, \tag{4.22}$$

intersection de la sphère unité avec le plan horizontal.

En parfaite analogie avec les propriétés du plan euclidien, on a alors :

Théorème 4.3 (Classification des isométries de la sphère S^2)**.** *Les isométries de la sphère unité sont exactement les restrictions à la sphère des rotations de l'espace ambiant de centre $O = (0, 0, 0)$, auxquelles il convient d'ajouter les réflexions dont le plan de référence passe par cette même origine, et les produits de rotations avec des réflexions.*

Lemme 4.2 (Principe de rétraction sur l'équateur)**.** *Si a et b sont deux points de l'équateur S^1 et si γ est un chemin joignant a à b dans la sphère S^2, alors la longueur de γ est minorée par celle de l'arc de cercle intermédiaire (que l'on peut calculer via la fonction arcsinus), avec égalité si et seulement si γ est astreint à se déplacer sur l'équateur, et prend "le bon côté" pour joindre a à b sur le cercle.*

Les preuves de ces lemmes obtenues (disons qu'il s'agit de bons exercices !), on n'a guère de mal à obtenir le célèbre résultat suivant :

Théorème 4.4 (Géodésiques de la sphère unité)**.** *Sur la sphère unité, les droites géodésiques sont exactement les* grands cercles*, ou intersections de la sphère S^2 avec un plan passant par l'origine O du repère.*

Entre deux points non-antipodaux de la sphère, il existe donc un unique plus court chemin.

Géométrie sphérique et axiomatique d'Euclide La sphère est un bon exemple d'espace "géométrique" – difficile de lui refuser ce qualificatif – qui n'obéit pas aux axiomes d'Euclide. Pour commencer, le Postulum d'Euclide est battu en brèche : étant donné une droite géodésique (i.e. un grand cercle) D et un point A extérieur à cette droite, impossible de trouver une nouvelle droite qui contienne A et qui n'intersecte pas D.

Mais, surtout, et au delà des questions axiomatiques, les propriétés classiques des triangles se retrouvent elles-mêmes toute chamboulées. Rappelez-vous : dans le plan, la somme des angles d'un triangle fait toujours 180°, trois angles droits font un rectangle et tout va pour le mieux dans le meilleur des mondes. Eh bien, comme tout bon explorateur (ou pilote de ligne) vous le dira : sur une sphère, c'est faux !

La Figure 4.5 en donne un bon exemple : on peut ainsi démontrer que « sur la sphère, la somme des angles d'un triangle fera toujours plus qu'un angle plat », et une quantité d'autres résultats qui, s'ils sont en opposition avec la théorie d'Euclide, conservent une cohérence interne impeccable.

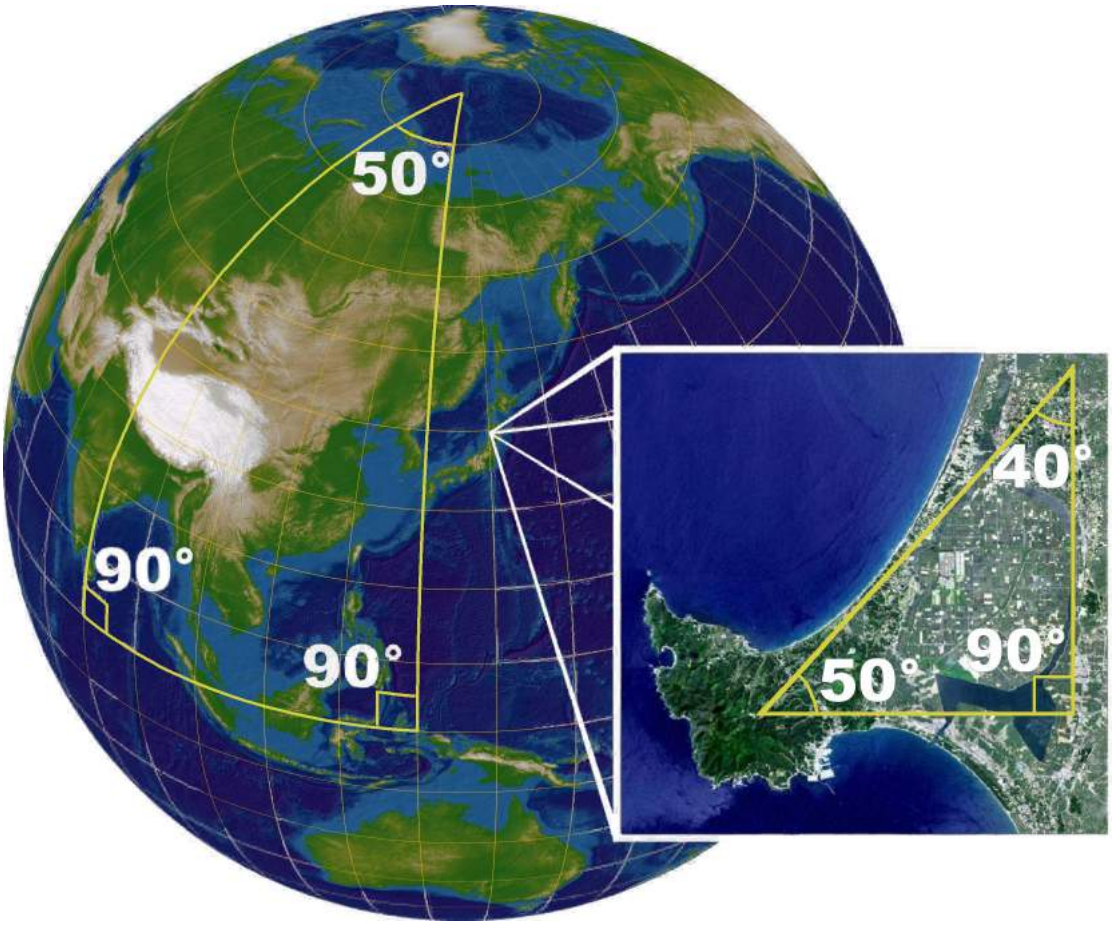


FIGURE 4.5 – Sur la sphère, la somme des angles d'un triangle géodésique est toujours strictement supérieure à 180° : c'est un espace de courbure positive. La géométrie euclidienne, *plate*, en donne néanmoins une bonne approximation pour des échelles très inférieures au rayon de courbure – dans le cas de la Terre, il est de 6 371km.
Image tirée de Wikipédia.

Métriques locales sur un ouvert de $\mathbb{R}^n$

La géométrie d'Euclide n'est donc pas la seule. Mais sera-t-il possible d'aller au delà des exemples canoniques du plan et de la sphère, d'envisager "toutes les géométries" possibles sur un monde de dimension 2 ? Il est impossible d'apporter une réponse décisive à une telle question "méta-mathématique" : qui pourrait en effet décider où s'arrête la "géométrie" ? Aux espaces homogènes ? Aux espaces continus ? Et l'île de Manhattan munie de la "distance des taxis", constitue-t-elle un espace "géométrique" ?

Plus une définition s'élargit, plus on gagne en généralité... Et moins on peut espérer de résultats forts, de descriptions fines. À chaque jeu d'hypothèses correspond donc une collection de théorèmes et un "domaine" plus ou moins pertinent : géométrie différentielle, algébrique, discrète... Pour le mathématicien, il s'agira avant tout de trouver des cadres d'étude réalisant un optimum entre simplicité conceptuelle, puissance des idées et généralité des applications : en bref, une théorie qui demande peu et donne beaucoup.

À ce petit jeu des théories, l'allemand Riemann s'est particulièrement distingué. C'est lui qui eut en 1854 l'idée de porter son attention sur les espaces qui sont topologiquement semblables à l'espace euclidien $\mathbb{R}^n$, mais qui sont munis d'une structure métrique *déformée*, *dilatée* par rapport à la distance euclidienne canonique.

> Supposons, par exemple, un monde renfermé dans une grande sphère et soumis aux lois suivantes :
>
> La température n'y est pas uniforme ; elle est maxima au centre, et elle diminue à mesure qu'on s'en éloigne, pour se réduire au zéro absolu quand on atteint la sphère où ce monde est renfermé.
>
> Je précise davantage la loi suivant laquelle varie cette température. Soit R le rayon de la sphère limite ; soit r la distance du point considéré au centre de cette sphère. La température absolue sera proportionnelle à $R^2 - r^2$.
>
> Je supposerai de plus que, dans ce monde, tous les corps aient même coefficient de dilatation, de telle façon que la longueur d'une règle quelconque soit proportionnelle à sa température absolue.
>
> Je supposerai enfin qu'un objet transporté d'un point à un autre, dont la température est différente, se met immédiatement en équilibre calorifique avec son nouveau milieu.
>
> Rien dans ces hypothèses n'est contradictoire ou inimaginable.
>
> Un objet mobile deviendra alors de plus en plus petit à mesure qu'on se rapprochera de la sphère limite.
>
> Observons d'abord que, si ce monde est limité au point de vue de notre géométrie habituelle, il paraîtra infini à ses habitants.
>
> Quand ceux-ci, en effet, veulent se rapprocher de la sphère limite, ils se refroidissent et deviennent de plus en plus petits. Les pas qu'ils font sont donc aussi de plus en plus petits, de sorte qu'ils ne peuvent jamais atteindre la sphère limite.
>
> Si, pour nous, la géométrie n'est que l'étude des lois suivant lesquelles se meuvent les solides invariables ; pour ces êtres imaginaires, ce sera l'étude des lois suivant lesquelles se meuvent les solides déformés par ces différences de température dont je viens de parler.
>
> Sans doute, dans notre monde, les solides naturels éprouvent également des variations de forme et de volume dues à l'échauffement ou au refroidissement. Mais nous négligeons ces variations en jetant les fondements de la géométrie ; car, outre qu'elles sont très faibles, elles sont irrégulières et nous paraissent par conséquent accidentelles.

Dans ce monde hypothétique, il n'en serait plus de même, et ces variations suivraient des lois régulières et très simples.

D'autre part, les diverses pièces solides dont se composerait le corps de ses habitants, subiraient les mêmes variations de forme et volume.

[...]

Ces êtres imaginaires seront donc comme nous conduits à classer les phénomènes dont ils seront témoins et à distinguer parmi eux, les « changements de position » susceptibles d'être corrigés par un mouvement volontaire corrélatif.

S'ils fondent une géométrie, ce ne sera pas comme la nôtre, l'étude des mouvements de nos solides invariables ; ce sera celle des changements de position qu'ils auront ainsi distingués, et qui ne sont autres que les « déplacements non euclidiens », ce sera la géométrie non euclidienne.

Ainsi des êtres comme nous, dont l'éducation se ferait dans un pareil monde, n'auraient pas la même géométrie que nous.

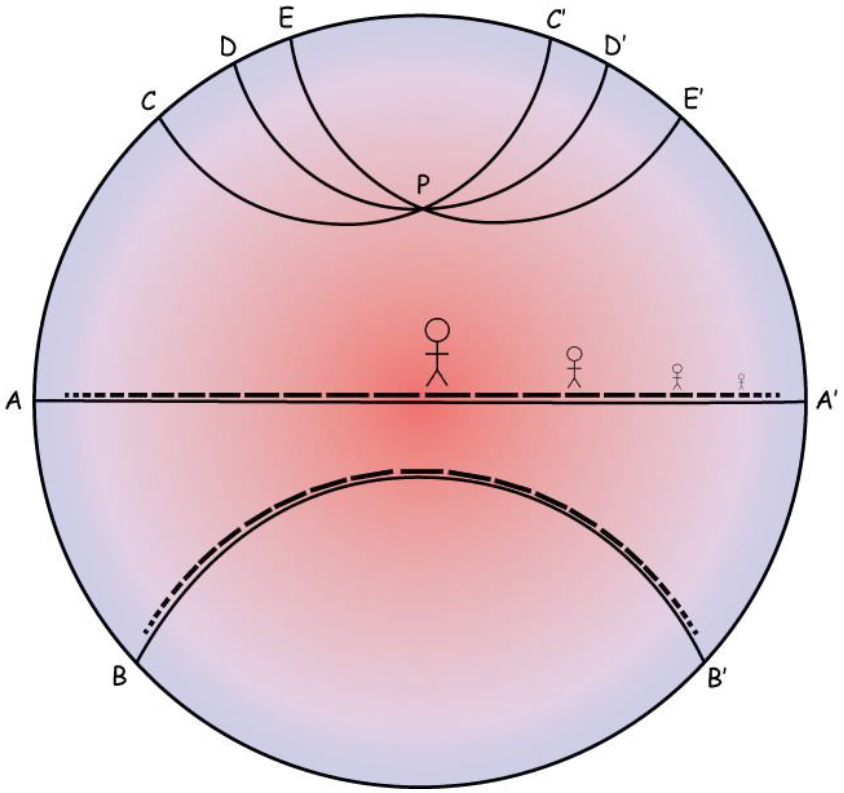


FIGURE 4.6 – Le disque de Poincaré. Pour emprunter le plus court chemin, un habitant de ce monde réduit aura toujours tendance à passer par le centre du disque, où la température élevée lui permet de faire des pas de géant. La simplicité de la métrique permet d'expliciter les "droites" géodésiques : ce sont les arcs de cercle orthogonaux au bord du disque. Par le point P passent donc une infinité de "droites" n'intersectant pas la droite (AA').
Image tirée et modifiée du site de John D. Norton, University of Pittsburgh.

Pour *construire* des géométries nouvelles, M. Poincaré propose donc simplement de partir d'un ouvert de l'espace euclidien – ici, la boule $\Omega = B(0,R)$ de $\mathbb{R}^2$ – et de le *déformer* localement : son champ de températures,

$$K(x) = R^2 - \|x\|^2 , \tag{4.23}$$

peut être compris comme un champ de dilatation/compression de l'espace.

Afin de s'adresser au sens physique de chacun, M. Poincaré s'encombre de plusieurs notions a priori indépendantes : température, coefficient de dilatation, indice de réfraction... Puisque nous avons ici à cœur de faire de la géométrie, seules les courbes, les distances, les volumes nous intéressent. Pour caractériser notre espace, on se contentera donc d'une formule donnant la longueur infinitésimale ressentie d'un petit segment $[(x,y),(x,y)+(\mathrm{d}x,\mathrm{d}y)]$ en tout point du disque :

$$\mathrm{d}_{\mathrm{poinc}}\Big((x,y)\to(x,y)+(\mathrm{d}x,\mathrm{d}y)\Big) = \frac{1}{R^2-\|(x,y)\|^2}\|(\mathrm{d}x,\mathrm{d}y)\| \tag{4.24}$$

ou, ce qui est plus pratique, sous la forme quadratique

$$\mathrm{d}^2_{\mathrm{poinc}}\Big((x,y)\to(x,y)+(\mathrm{d}x,\mathrm{d}y)\Big) = \frac{1}{\big(R^2-(x^2+y^2)\big)^2}\big(\mathrm{d}x^2+\mathrm{d}y^2\big) . \tag{4.25}$$

Cette formule est localement équivalente au théorème de Pythagore de l'espace euclidien,

$$\mathrm{d}^2_{\mathrm{eucl}}\Big((x,y)\to(x,y)+(\mathrm{d}x,\mathrm{d}y)\Big) = \big(\mathrm{d}x^2+\mathrm{d}y^2\big) , \tag{4.26}$$

avec une différence de taille : la présence d'un coefficient de dilatation $1/K(x)^2$, qui déforme la métrique et rend les longueurs infinies au bord du disque – i.e. lorsque $x^2+y^2=R^2$.

Longueurs de courbes Étant donné une forme quadratique

$$\mathrm{d}^2_{(x,y)} : (\mathrm{d}x,\mathrm{d}y) \mapsto d^2\Big((x,y)\to(x+\mathrm{d}x,y+\mathrm{d}y)\Big), \tag{4.27}$$

que l'on appellera *métrique Riemannienne locale* au point x, on peut maintenant définir la longueur d'un chemin $\gamma : [0,1]\to\Omega$ à valeur dans notre domaine comme la somme des longueurs des pas infinitésimaux $\mathrm{d}\gamma(t) = \dot\gamma(t)\,\mathrm{d}t$, calculées au sens de la métrique locale $\mathrm{d}^2_{\gamma(t)}$:

$$\ell(\gamma) = \int_0^1 \sqrt{\mathrm{d}^2(\gamma(t)\to\gamma(t)+\dot\gamma(t)\,\mathrm{d}t)} \tag{4.28}$$

$$= \int_0^1 \sqrt{\mathrm{d}^2_{\gamma(t)}(\dot\gamma(t))}\,\mathrm{d}t. \tag{4.29}$$

Il s'agit d'une simple généralisation aux métriques "non-uniformes" de la formule (4.7) donnant la longueur euclidienne d'une courbe, et on peut par exemple calculer le "rayon" du disque de Poincaré associé au chemin radial $\gamma(t) = (t,0)$. Pour la métrique euclidienne, habituelle, on a sans surprise :

$$\ell_{\mathrm{eucl}}(\gamma) = \int_0^1 \|\dot\gamma(t)\|_{\gamma(t),\ \mathrm{eucl}}\,\mathrm{d}t \tag{4.30}$$

$$= \int_0^1 \|(1,0)\|_{\gamma(t),\ \mathrm{eucl}}\,\mathrm{d}t \tag{4.31}$$

$$= \int_0^1 \sqrt{1^2+0^2}\ \mathrm{d}t \qquad = \quad 1. \tag{4.32}$$

Pour la métrique de Poincaré, on a par contre :

$$\ell_{\text{poinc}}(\gamma) = \int_0^1 \|\dot\gamma(t)\|_{\gamma(t),\ \text{poinc}}\,\mathrm{d}t \tag{4.33}$$

$$= \int_0^1 \|(1,0)\|_{\gamma(t),\ \text{poinc}}\,\mathrm{d}t \tag{4.34}$$

$$= \int_0^1 \frac{\|(1,0)\|_{\gamma(t),\ \text{eucl}}}{1-\|\gamma(t)\|^2}\,\mathrm{d}t \tag{4.35}$$

$$= \int_0^1 \frac{\sqrt{1^2+0^2}}{1-t^2}\,\mathrm{d}t \tag{4.36}$$

$$= \int_0^1 \frac{1}{1+t}\cdot\frac{1}{1-t}\,\mathrm{d}t \quad = \quad +\infty \tag{4.37}$$

Le disque de Poincaré est donc un espace métrique *non borné* : un univers dans une coquille de noix. Son étude explicite fera l'objet des pages suivantes, et nous verrons que le Postulum d'Euclide n'y est pas vérifié – la figure 4.6 permet déjà de s'en faire une bonne intuition.

Géodésiques du disque de Poincaré

Au chapitre 6, nous étudierons les applications concrètes de la géométrie Riemannienne à l'anatomie et au traitement d'images médicales. Mais aujourd'hui, plutôt que d'entamer un long discours de vulgarisation, je voudrais vous faire vibrer un peu. Vous présenter un exemple de *belle* preuve, le genre d'images qui laisse un mathématicien songeur.

Il s'agira de démontrer le fait suivant sur les géodésiques (ou plus courts chemins) du monde hyperbolique décrit précédemment :

Théorème 4.5 (Description des géodésiques du disque de Poincaré)**.** *Les plus courts chemins du disque de Poincaré sont exactement :*

- *les arcs de cercles orthogonaux aux bords du disque ;*
- *les diamètres droits, qui passent par son centre.*

On notera que les seconds peuvent être vus simplement comme des arcs de cercle de rayon infini.

Entre deux points donnés du disque, il existe donc un unique plus court chemin. Mais, conformément à ce qui est énoncé Figure 4.6, le disque de Poincaré ne vérifie pas le Postulum d'Euclide.

On pourrait démontrer ce résultat par le calcul : la forme particulièrement simple de la métrique donnée équation 4.25 permet en effet de conserver des développements raisonnables. Mais aujourd'hui, je voudrais vous présenter une preuve entièrement géométrique, due au grand géomètre William Thurston (1946-2012), et rapportée dans l'excellent article d'introduction *Hyperbolic Geometry*, de J.W. Cannon et al., auquel je réfère le lecteur avide de précisions et de rigueur technique.

Projections et changements de coordonnées

Dans cette preuve, nous utiliserons de manière intensive les *symétries* cachées de la métrique de Poincaré. Pour les mettre en évidence, nous utiliserons de nombreuses *projections* du disque vers un hémisphère, un demi-plan.

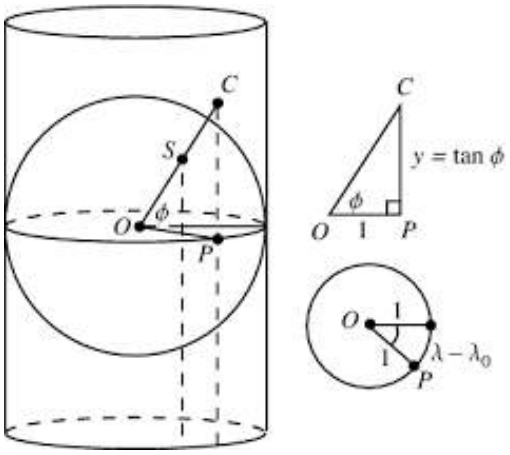
Une fois n'est pas coutume, pas de grande nouveauté ici : les projections des mathématiciens sont bien celles employées de tout temps par les géographes pour représenter sur un plan la géométrie sphérique de notre globe terrestre.

Comme pour la projection de Mercator présentée Figure 4.7, on prendra toutefois garde à bien considérer sur l'espace d'arrivée la métrique induite par l'espace de départ : sur un planisphère "déformé" comme sur la Terre, le Groenland n'est pas plus vaste que l'Afrique tout entière !

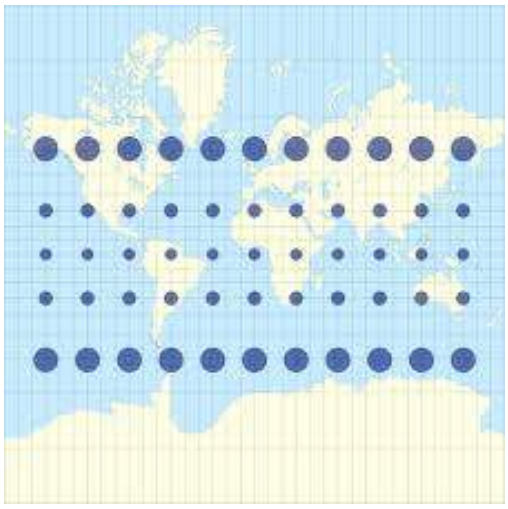
Plutôt que des projections de type Mercator, nous préférerons ici des projections *stéréographiques*, qui conservent les angles – voir Figure 4.8. Comme démontré à la fin du film *Dimensions* d'Étienne Ghys, Jos Leys et Aurélien Alvarez, ces projections ont de nombreuses propriétés de conservation, toutes très utiles d'un point de vue théorique.

D'abord, elles conservent les angles : si deux courbes γ_1, γ_2 sur la sphère se croisent en p avec un angle θ, alors les courbes projetées $F \circ \gamma_1$ et $F \circ \gamma_2$ se croisent en $F(p)$ avec ce même angle θ. Ensuite, elles préservent localement les rapports des surfaces : localement, la carte au point $F(p)$ est identique à celle au point p, à un facteur de dilatation près. Enfin, et c'est cette dernière propriété qui sera à retenir pour la suite : les projections stéréographiques envoient des cercles de la sphère sur des cercles (ou droites) du plan, comme illustré Figure 4.9.

Pour bien nous approprier ces notions, nous projetterons en classe le Chapitre 1 du film Dimensions, dont sont tirées les belles images ci-contre – pour les démonstrations, je vous invite à regarder le Chapitre 9, librement accessible à l'adresse suivante : `www.dimensions-math.org`.



(a) La projection de Mercator : $C = M(S)$



(b) Les cercles bleus sont de rayons constant pour la métrique induite de la sphère sur le plan.

FIGURE 4.7 – Projection de Mercator et métrique induite sur le plan.
Images tirées de Wikipédia.

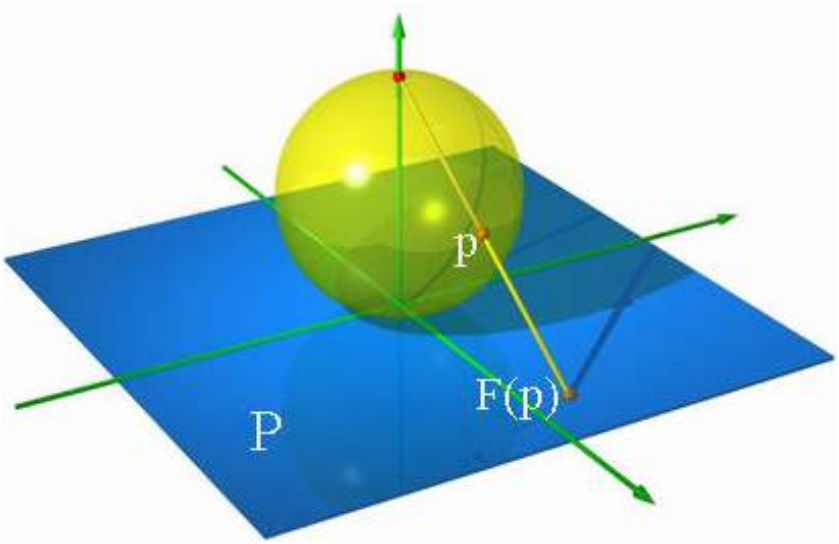


FIGURE 4.8 – Projection stéréographique de la sphère sur le plan tangent au pôle Sud.

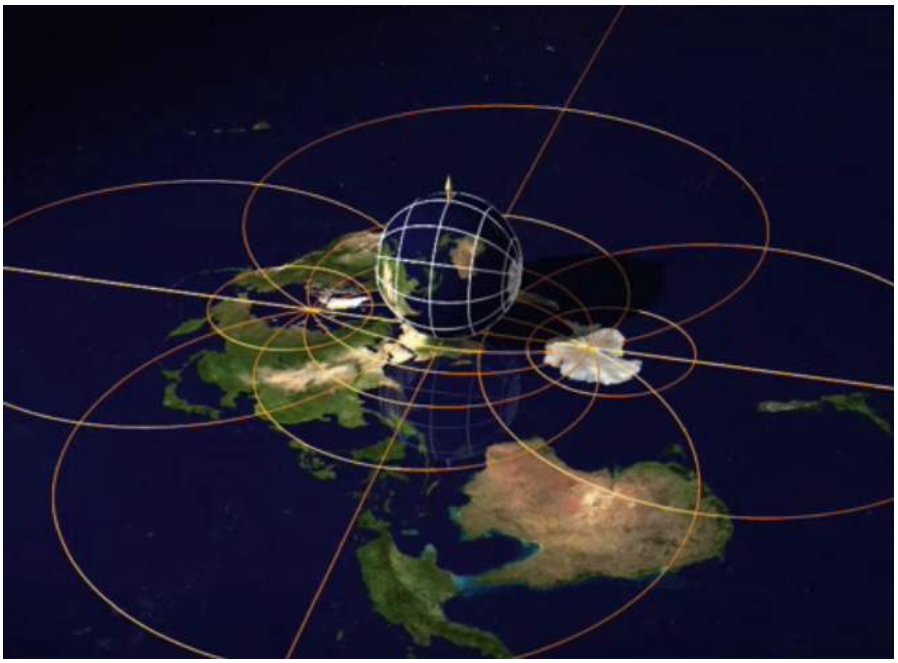


FIGURE 4.9 – Une projection stéréographique envoie les cercles sur des droites ou des cercles.

Modèles standards de la géométrie hyperbolique

Nous avons vu précédemment comment déformer le disque unité I pour le rendre conforme aux idées de Poincaré. On propose maintenant de le plonger dans l'espace ambiant $\mathbb{R}^3$, en écrivant :

$$I = \{(x_1, x_2, 0) \mid x_1^2 + x_2^2 < 1\} \tag{4.38}$$

muni de la métrique :

$$\mathrm{d}s_I^2 = 4\frac{\mathrm{d}x_1^2 + \mathrm{d}x_2^2}{(1 - x_1^2 - x_2^2)^2}. \tag{4.39}$$

On a simplement remplacé x par x_1, et y par x_2, en plus de dilater la métrique par un facteur 4 – pour des questions de normalisation de la courbure, qui importent peu ici. Suivant la Figure 4.10, on considère deux modèles supplémentaires : l'hémisphère

$$J = \{(x_1, x_2, x_3) \mid x_1^2 + x_2^2 + x_3^2 = 1 \text{ et } x_3 > 0\} \tag{4.40}$$

et le demi-plan

$$H = \{(1, x_2, x_3) \mid x_3 > 0\}. \tag{4.41}$$

Pour passer de I à J puis de J à H, on définit les deux projections stéréographiques β et α, qui sont respectivement associées aux points $(0, 0, -1)$ et $(-1, 0, 0)$. Un point courant i du disque est donc associé de manière unique à un point $j = \beta(i)$ de l'hémisphère, puis à un point $h = \alpha(j)$ du demi-plan.

A priori, toutes ces projections, ces changements de coordonnées ne nous mènent pas bien loin : certes, le disque de Poincaré peut être vu au travers de nos projections – de nos miroirs déformants – comme un hémisphère, un demi-plan... Mais encore ?

On peut déjà constater que, par les propriétés des projections stéréographiques énoncées plus haut, un cercle sur I correspond exactement à un cercle sur J, puis à un cercle sur H. Plus exactement, on a l'équivalence suivante :

Lemme 4.3 (Grands cercles sur les modèles du plan hyperbolique)**.** *Soit $\gamma : \mathbb{R} \to I$ un chemin, une courbe définie à valeur dans le disque de Poincaré. Alors les trois propositions suivantes sont équivalentes :*

1. *γ décrit un arc de cercle orthogonal aux bords du disque, ou bien un diamètre de ce dernier.*
2. *$\beta \circ \gamma$, qui est un chemin à valeurs dans J, décrit un demi-cercle sur J qui coupe de manière orthogonale le plan horizontal.*
3. *$\alpha \circ \beta \circ \gamma$, qui est un chemin à valeurs dans H, décrit un demi-cercle qui coupe de manière orthogonale l'axe horizontal, ou bien est une droite verticale.*

Démonstration. C'est une conséquence directe des propriétés énoncées plus haut. Je vous conseille néanmoins de faire de nombreux croquis dans la marge du présent polycopié, pour bien visualiser les chemins en question et vous convaincre de la véracité de mes propos. Il va sans dire que je ferai quelques dessins en classe ! □

Ces équivalences entre "cercles généralisés" sont un bon début. Mais le véritable intérêt de notre manœuvre réside dans la proposition suivante, qui décrit complètement la métrique induite du disque de Poincaré I vers l'hémisphère J et le demi-plan H.

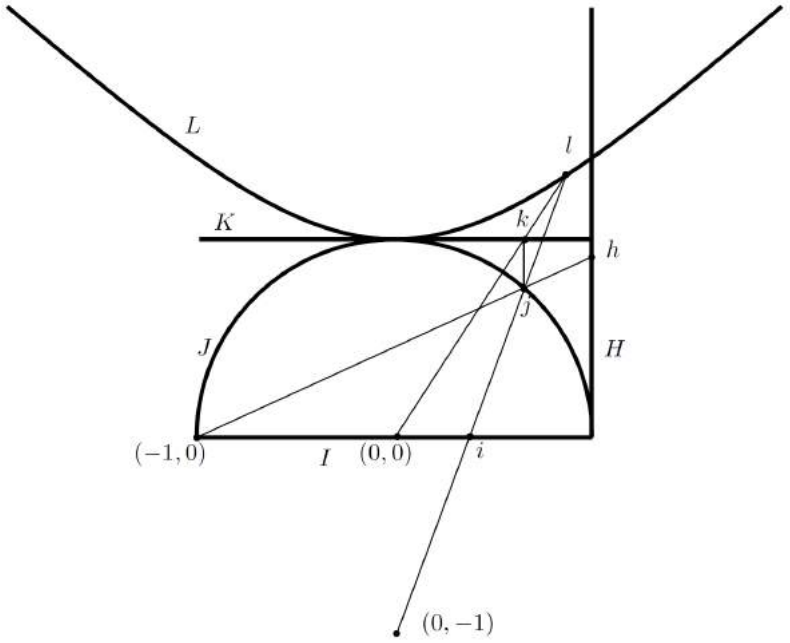


FIGURE 4.10 – Les cinq modèles du plan hyperbolique vus dans le plan (x_1, x_3), c'est à dire en coupe de côté. Les projections associées transportent la métrique du disque I vers l'hémisphère J, puis de là vers le demi-plan H. Les deux derniers modèles K et L sont plus anecdotiques, et ne seront pas abordés ici.
Dessin tiré de l'article de J. W. Cannon et al., *Hyperbolic Geometry.*

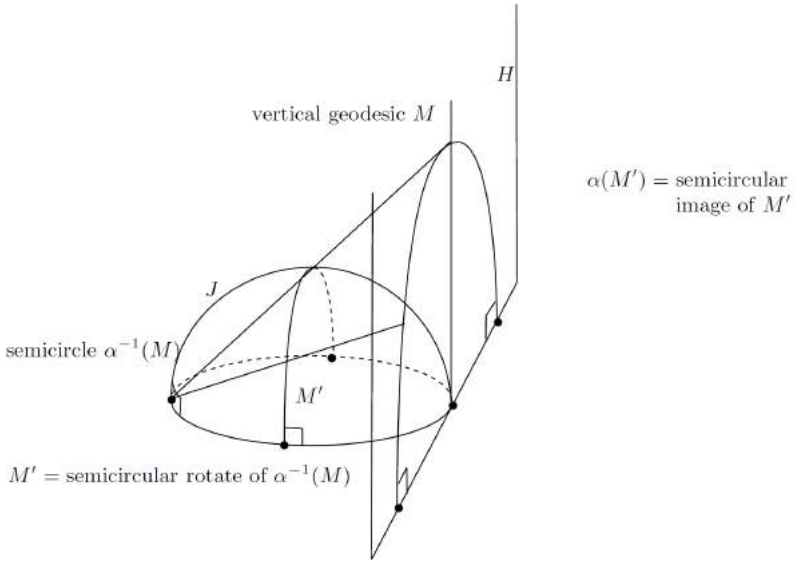


FIGURE 4.11 – Selon l'orientation, un demi-grand cercle de J peut-être envoyé par α sur un demi-cercle de H, ou sur l'axe vertical.
Dessin tiré de l'article de J. W. Cannon et al., *Hyperbolic Geometry.*

Proposition 4.1. *De même que la projection de Mercator permettait de munir le plan d'une métrique induite le rendant isométrique au globe terrestre privé de ses pôles, les projections stéréographiques β et α permettent de munir l'hémisphère J et le demi-plan H de métriques riemanniennes (i.e. de "champs de température") les rendant isométriques au disque de Poincaré.*

Elles sont données par les formules suivantes :

$$\mathrm{d}s_J^2 = \frac{\mathrm{d}x_1^2 + \mathrm{d}x_2^2 + \mathrm{d}x_3^2}{x_3^2} \tag{4.42}$$

$$\mathrm{d}s_H^2 = \frac{\mathrm{d}x_2^2 + \mathrm{d}x_3^2}{x_3^2}, \tag{4.43}$$

qu'il est bon de mettre en parallèle avec la métrique du disque,

$$\mathrm{d}s_I^2 = 4\frac{\mathrm{d}x_1^2 + \mathrm{d}x_2^2}{(1 - x_1^2 - x_2^2)^2}. \tag{4.44}$$

Autrement dit, là où Poincaré munissait le disque d'un champ de température

$$K_I(x_1, x_2) = \frac{1}{2}\Big(1 - (x_1^2 + x_2^2)\Big), \tag{4.45}$$

on peut munir l'hémisphère et le demi-plan de champs de température analogues,

$$K_J(x_1, x_2, x_2) = x_3 \tag{4.46}$$

$$K_H(x_2, x_3) = x_3. \tag{4.47}$$

Enfin, si γ est un chemin sur I, et en prenant $\beta \circ \gamma$, $\alpha \circ \beta \circ \gamma$ les chemins images sur J et H, alors, au sens des températures/métriques définies ci-dessus, on a

$$l_I(\gamma) = l_J(\beta \circ \gamma) = l_H(\alpha \circ \beta \circ \gamma). \tag{4.48}$$

La démonstration de cette proposition repose sur un calcul qui, s'il est immédiat, me semble un peu trop technique pour des élèves de filières non scientifiques. On peut tout de même essayer de le comprendre !

Commençons par remarquer que, dans le modèle de l'hémisphère comme dans le domaine du demi-plan, la température tend vers 0 lorsqu'on s'approche du plan horizontal, du bord du domaine. Ainsi (on peut le vérifier comme à l'équation (4.37)), le bord reste toujours à l'infini : dans H et J comme dans I, un chemin qui sort du domaine est nécessairement de longueur infinie.

Plus que cette simple vérification, ce qui nous intéresse ici est la forme bien particulière des champs de températures K_J et K_H : tous deux dépendent uniquement de x_3 de façon linéaire. Autrement dit, et c'est un point crucial, la métrique est invariante par toute une famille de transformations du domaine faciles à comprendre :

Théorème 4.6 (Premières isométries du plan hyperbolique)**.** *On peut caractériser quatre familles d'isométries du plan hyperbolique :*

1. *Les réflexions de l'hémisphère J telles que*

$$\rho(x_1, x_2, x_3) = (-x_1, x_2, x_3). \tag{4.49}$$

2. *Les rotations R_θ de l'hémisphère J autour de l'axe (Oz), qui conservent les lignes de niveau horizontales :*

$$R_\theta(x_1, x_2, x_3) = (x_1\cos(\theta) + x_2\sin(\theta), -x_1\sin(\theta) + x_2\cos(\theta), x_3) \tag{4.50}$$

3. *Les translations horizontales τ_x du demi-plan H, qui conservent elles aussi la composante d'altitude :*

$$\tau_x(1, x_2, x_3) = (1, x_2 + x, x_3). \tag{4.51}$$

4. *Les dilatations σ_h du demi-plan H, pour toute valeur du paramètre $h > 0$:*

$$\sigma_h(1, x_2, x_3) = (1, h \cdot x_2, h \cdot x_3). \tag{4.52}$$

Démonstration. Les trois premiers points sont immédiats et sans surprise : ils reposent simplement sur le fait que réflexions, rotations et translations horizontales conservent à la fois la norme des vecteurs vitesses et la composante d'altitude. Prouvons par exemple le troisième, et prenons un chemin $\gamma : [0,1] \to H$. À chaque instant t de l'intervalle $[0,1]$, γ associe un triplet $(1, \gamma_2(t), \gamma_3(t))$ appartenant au demi-plan H – on a donc $\gamma_3(t) > 0$. Pour tout x réel, le translaté horizontal $\tau_x \circ \gamma$ est simplement donné par :

$$\forall\, t \in [0,1], \tau_x \circ \gamma(t) = (1, \gamma_2(t) + x, \gamma_3(t)), \tag{4.53}$$

$$\dot{\widetilde{\tau_x \circ \gamma}}(t) = (0, \dot{\gamma}_2(t) + 0, \dot{\gamma}_3(t)). \tag{4.54}$$

On a alors :

$$\begin{aligned}
l_H(\tau_x \circ \gamma) &= \int_0^1 \left\| \dot{\widetilde{\tau_x \circ \gamma}}(t) \right\|_{\tau_x \circ \gamma(t)} \mathrm{d}t && (4.55)\\
&= \int_0^1 \|(0, \dot{\gamma}_2(t) + 0, \dot{\gamma}_3(t))\|_{(1, \gamma_2(t)+x, \gamma_3(t))}\, \mathrm{d}t && (4.56)\\
&= \int_0^1 \frac{\sqrt{\dot{\gamma}_2(t)^2 + \dot{\gamma}_3(t)^2}}{\gamma_3(t)} \mathrm{d}t && (4.57)\\
&= \int_0^1 \|(0, \dot{\gamma}_2(t), \dot{\gamma}_3(t))\|_{(1, \gamma_2(t), \gamma_3(t))}\, \mathrm{d}t && (4.58)\\
&= \int_0^1 \|\dot{\gamma}(t)\|_{\gamma(t)}\, \mathrm{d}t && (4.59)\\
&= l_H(\gamma) && (4.60)
\end{aligned}$$

La translation horizontale τ_x conserve donc bien la longueur des chemins : c'est une isométrie.

Les rotations, réflexions et translations *hyperboliques* que nous avons exhibé ici sont exactement les analogues des rotations, réflexions et translations "classiques" du plan euclidien. Ce sont des transformations simples, qui préservent les longueurs des chemins, et caractérisent une certaine *homogénéité* de l'espace métrique étudié. Que les dilatations σ_h du demi-plan soient

elles-aussi des isométries du "plan hyperbolique" est, par contre une réelle surprise : dans le plan euclidien, les dilatations modifient les échelles sans conserver les longueurs. En un sens, nous allons donc prouver que *le groupe des isométries du plan hyperbolique est plus riche que celui des isométries du plan euclidien.* Fixons une valeur du paramètre h qui soit strictement positive. On a alors

$$\forall\, t \in [0,1], \sigma_h \circ \gamma(t) = (1, h \cdot \gamma_2(t), h \cdot \gamma_3(t)) \tag{4.61}$$

$$\dot{\widetilde{\sigma_h \circ \gamma}}(t) = (0, h \cdot \dot{\gamma}_2(t), h \cdot \dot{\gamma}_3(t)). \tag{4.62}$$

On peut calculer la longueur du chemin dilaté :

$$l_H(\sigma_h \circ \gamma) = \int_0^1 \left\| \dot{\widetilde{\sigma_h \circ \gamma}}(t) \right\|_{\sigma_h \circ \gamma(t)} \mathrm{d}t \tag{4.63}$$

$$= \int_0^1 \|(0, h \cdot \dot{\gamma}_2(t), h \cdot \dot{\gamma}_3(t))\|_{(1, h\cdot\gamma_2(t), h\cdot\gamma_3(t))} \,\mathrm{d}t \tag{4.64}$$

$$= \int_0^1 \frac{\sqrt{h^2 \cdot \dot{\gamma}_2(t)^2 + h^2 \cdot \dot{\gamma}_3(t)^2}}{h \cdot \gamma_3(t)} \mathrm{d}t \tag{4.65}$$

$$= \int_0^1 \|(0, \dot{\gamma}_2(t), \dot{\gamma}_3(t))\|_{(1, \gamma_2(t), \gamma_3(t))} \,\mathrm{d}t \tag{4.66}$$

$$= \int_0^1 \|\dot{\gamma}(t)\|_{\gamma(t)} \,\mathrm{d}t \tag{4.67}$$

$$= l_H(\gamma). \tag{4.68}$$

Comme attendu, la dilatation σ_h conserve bien la longueur des chemins. □

À l'aide de nos projections stéréographiques, voilà que, sans grand effort, nous avons obtenu un théorème de "classification" des isométries du plan hyperbolique. Celui-ci nous sera, on s'en doute, d'un grand secours pour déterminer les géodésiques de notre espace. Reste simplement à trouver un principe de rétraction adéquat : il est donné par le lemme ci-dessous.

Lemme 4.4 (Principe de rétraction dans le plan hyperbolique)**.** *Soit $a = (1, 0, 2)$ et $b = (1, 0, x)$ deux points du demi-plan hyperbolique H,* ***situés sur la droite verticale "centrale"****, qui touche I et J en sa limite inférieure. On définit à cette occasion le segment hyperbolique*

$$s_{a\to b} : t \in [0,1] \mapsto (1, 0, (1-t)\cdot 2 + t \cdot x), \tag{4.69}$$

(qui n'est pas de vitesse constante au sens de la métrique sur H). Alors pour tout chemin γ joignant a à b dans le demi-plan H, on a

$$\ell_H(\gamma) \;\geqslant\; \ell_H(s_{a\to b}) \;=\; \int_0^1 \frac{x}{2 + t\cdot(x-2)} \mathrm{d}t = \frac{x \log(x/2)}{x-2}, \tag{4.70}$$

avec cas d'égalité si et seulement si γ est égal au segment à $s_{a\to b}$ à reparamétrisation près.

Démonstration. La démonstration du lemme suit dans les grandes lignes celle que nous avions développé dans le cas euclidien, avec décomposition en une partie "utile" (dans notre cas, verticale), et partie "inutile" (ici, horizontale). Les détails sont laissés à la sagacité du lecteur. □

Tous les outils en main, on peut maintenant attaquer le cœur de notre séance :

Preuve du théorème de classification des géodésiques du disque de Poincaré. On travaillera dans le demi-plan H, avant d'utiliser le Lemme 4.3 pour transporter nos résultats sur le disque.

On se donne donc deux points distincts $a = (1, a_2, a_3)$ et $b = (1, b_2, b_3)$ du demi-plan, et on définit le segment hyperbolique $s_{a\to b} : [0, 1] \to H$ comme suit :

— Si $a_2 = b_2$, i.e. si a et b sont sur la même droite verticale, alors $\sigma_{a\to b}$ est simplement le segment vertical défini dans la page précédente.

— Sinon, c'est qu'il existe un unique cercle $C_{a,b}$ du plan "$x_1 = 1$" qui contienne à la fois a et b, et dont le centre soit situé sur l'axe $x_3 = 0$; $C_{a,b}$ est donc un cercle qui coupe le bord de H de façon orthogonale. On peut en effet remarquer que le seul centre convenable pour $C_{a,b}$ est à l'intersection entre le bord de H et la médiatrice du segment euclidien $[a, b]$. On définit alors $s_{a\to b}$ comme l'arc de $C_{a,b}$ allant de a à b à vitesse euclidienne constante.

Étant donné un chemin γ quelconque joignant a à b dans H, **il s'agit de montrer que :**

$$\ell_H(\gamma) \;\geqslant\; \ell_H(s_{a\to b}) \tag{4.71}$$

avec cas d'égalité si et seulement si $\gamma = s_{a\to b}$ à reparamétrisation près.

On travaille par étapes :

— Tout d'abord, on utilise **une translation et une dilatation** de H pour ramener a sur le point $(1, 0, 2)$: avec $f = \sigma_{2/a_3} \circ \tau_{-a_2}$, on a

$$f(a) = (1, 0, 2), \qquad \ell_H(f \circ \gamma) = \ell_H(\gamma), \qquad f \circ s_{a\to b} = s_{(1,0,2)\to f(b)}. \tag{4.72}$$

— Si $f(b)$ est sur la droite $b_2 = 0$, alors c'est gagné, d'après le lemme de rétraction. Sinon, on utilise la **projection stéréographique** α conformément à la Figure 4.11. Comme $f(a) = (1, 0, 2) = \alpha((0, 0, 1))$, on sait que $\alpha^{-1}(f(a))$ correspond au pôle Nord de l'hémisphère J. L'image réciproque de l'arc de cercle $f \circ s_{a\to b}$ est donc exactement l'arc de "grand cercle" partant du pôle Nord de J pour aller au point $\alpha^{-1}(f(b))$.

De son côté, le chemin $f \circ \gamma$ est transporté vers un chemin a priori quelconque reliant le pôle Nord à $\alpha^{-1}(f(b))$, avec toutefois

$$\ell_J(\alpha^{-1} \circ f \circ \gamma) = \ell_H(f \circ \gamma) = \ell_H(\gamma), \tag{4.73}$$

puisque les métriques sur J et H se correspondent exactement via α.

— À l'aide d'une **rotation** R_θ de l'hémisphère J autour de l'axe vertical, on positionne l'image de b sur le quart de cercle de J donné par " $x_1 > 0$, $x_2 = 0$, $x_3 > 0$ ".

— Enfin, à l'aide de α, on **re-projette le tout sur** H. Bilan total de notre opération : nous disposons d'une isométrie $g = \alpha \circ R_\theta \circ \alpha^{-1} \circ \sigma_{2/a_3} \circ \tau_{-a_2}$ de H telle que :

$$g(a) = (1, 0, 2), \quad \ell_H(g \circ \gamma) = \ell_H(\gamma), \quad \text{et} \quad g(b) \text{ sur l'axe médian "} x_2 = 0 \text{" de } H. \tag{4.74}$$

Surtout, comme les translations, dilatations, rotations et projections stéréographiques conservent la propriété d'être un arc de cercle normal aux bords du domaine, on a que $g \circ s_{a\to b} = s_{(1,0,2)\to g(b)}$ – à la paramétrisation près.

Appliquer le principe de rétraction aux points $g(a)$ et $g(b)$ nous donne alors exactement le résultat voulu, à savoir que l'optimum de longueur est atteint si et seulement si on a $g \circ \gamma = s_{(1,0,2)\to g(b)} = g \circ s_{a\to b}$, i.e. $\gamma = s_{a\to b}$ à paramétrisation près ; Cqfd.

□

Bilan

À la première lecture, je ne doute pas que les pages précédentes vous aient laissés perplexes : tant de remarques, de lemmes, de résultats annexes pour un si petit résultat ! C'est que, pour une fois, j'ai souhaité vous montrer une véritable preuve mathématique jusque dans ses complications techniques – en écrémant pourtant la partie relative aux problèmes de paramétrisation. Tous ces détails sont là pour vous rassurer, vous servir de référence... Pas pour obscurcir votre vue ! Pour clore cette section, je souhaiterais donc mettre en avant les idées essentielles qui nous ont permis de résoudre le problème des géodésiques du disque de Poincaré de manière efficace.

Première entre toutes, la place de choix accordée aux isométries. Que ce soit dans le cas du plan euclidien, de la sphère ou du plan hyperbolique, nous avons su tirer parti de l'*homogénéité* des espaces étudiés, de la richesse de leur groupe d'isométries : pour tout couple de points a et b, nous avons réussi à expliciter une isométrie qui envoyait a sur un point de référence (l'origine, un point de l'équateur, le point $(1, 0, 2)$), et b sur un espace favorable au calcul (l'axe horizontal, l'équateur, l'axe médian de H).

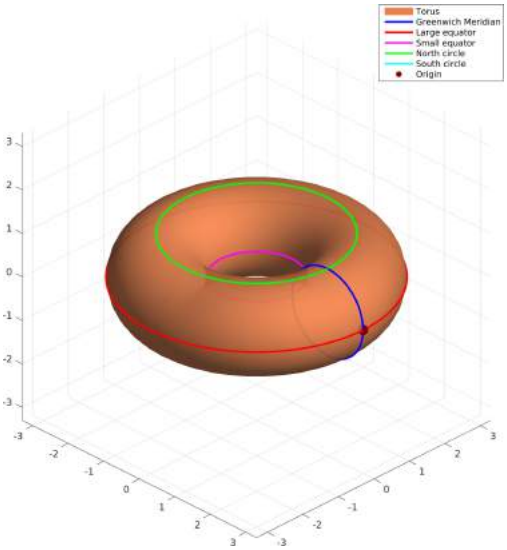
L'attention accordée à une description simple du groupe des isométries. Pour cela, il était indispensable d'obtenir une décomposition pratique des isométries en classes primitives élémentaires. Si l'habitude nous a permis d'obtenir des résultats rapides dans le cas du plan euclidien et de la sphère, il nous a fallu déployer des trésors d'ingéniosité pour décrire le groupe, bien plus riche, des isométries du plan hyperbolique. L'introduction des **projections stéréographiques** α, β et des **trois modèles standards** I, J et H nous a heureusement permis de conserver une certaine intuition, et d'assimiler les isométries élémentaires R_θ, τ_x et σ_h à des opérations "courantes" de l'espace euclidien.

Le recours au calcul infinitésimal pour montrer un principe de réduction atomique. Toutes ces transformations, ces isométries avaient pour but de nous ramener à un cadre normalisé, dans lequel les calculs étaient simples... Restait encore à les effectuer : une fois de plus, c'est le calcul infinitésimal de Leibniz et Newton qui nous a tiré d'affaire, en donnant le formalisme adéquat pour comparer et découper des longueurs de chemins arbitrairement compliqués.

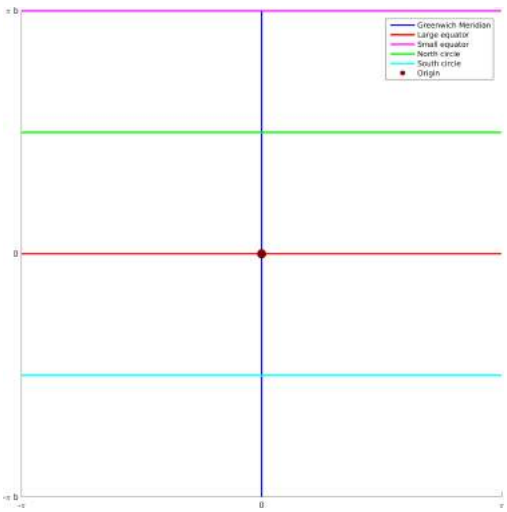
En fin de compte, nous avons donc obtenu une preuve intelligible, bien découpée, avec un recours minimal au calcul : un bien beau résultat aux yeux du géomètre, qui tient maintenant un exemple explicite et bien compris d'espace ne vérifiant pas le Postulum d'Euclide.

Intérêt de la géométrie Riemannienne : l'exemple du tore

Déformer judicieusement l'espace euclidien permet donc de construire des *modèles* aux théories axiomatiques non-euclidiennes développées par Lobatchevski et Bolyai. L'idée clé de Riemann aura été de généraliser cette procédure à *tous* les étirements réguliers de la métrique euclidienne sur $\mathbb{R}^n$: de manière encore plus générale qu'avec les seuls champs de températures, on s'autorisera à considérer tout champ de matrices $n \times n$ symétriques définies positives, $g : x \in \Omega \mapsto g(x) \in S_n^{++}$ pour pénaliser les petits déplacements au point x, avec g de classe C^∞ et Ω un ouvert de $\mathbb{R}^n$ – ou mieux, une variété de dimension n, c'est à dire un espace localement (mais pas nécessairement globalement) isomorphe à $\mathbb{R}^n$.

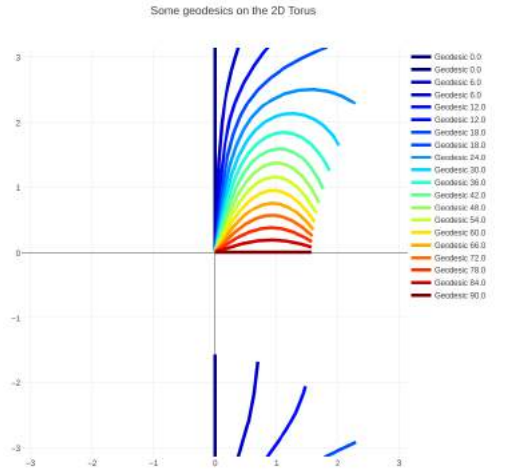


(a) Un tore *bouée* vu comme surface immergée dans l'espace ambiant $\mathbb{R}^3$.

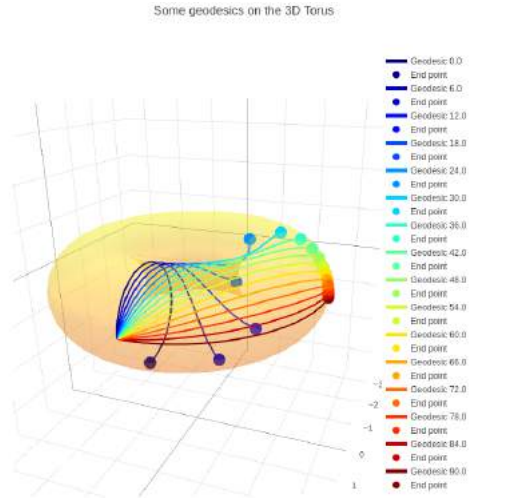


(b) Le tore carré. Muni de la métrique naturelle euclidienne, il est *plat* : petits et grands parallèles sont de même longueur.

FIGURE 4.12 – Deux manières de représenter un tore. La géométrie Riemannienne permet d'étudier la structure métrique de la bouée avec le confort pratique offert par une paramétrisation bidimensionnelle.



(a) Par contre, en déformant ce rectangle par une métrique bien choisie (qui pénalise certaines droites horizontales plus que d'autres, selon qu'elles correspondent à des grands ou à de petits équateurs), on peut obtenir une structure géométrique identique à celle du tore bouée.



(b) Les courbes géodésiques, tirées numériquement sur la représentation de gauche (sans avoir à se soucier d'un quelconque "décollement") donnent bien les résultats attendu en 3D !

FIGURE 4.13 – Géodésiques de longueur $3\pi/2$ tirées à partir du grand équateur, sur le tore de rayon médian 2 et de section 1. Les trajectoires sont intégrées par une simple méthode d'Euler, sur un domaine rectangulaire périodique $(\theta, \chi) \in S^1 \times S^1$.

L'idée est féconde : prenons l'exemple du tore. Il existe a priori deux manières de l'étudier : en tant que surface immergée dans l'espace ambiant $\mathbb{R}^3$, ou comme quotient du plan $\mathbb{R}^2$ par l'action des translations entières – le fameux carré aux bords recollés. Cette dernière représentation est de loin la plus pratique : plutôt que de travailler avec des points (x, y, z) dont il faut toujours s'assurer qu'ils ne quittent pas la surface, on utilise des couples (r, θ) sans se poser d'autre question que celle du "modulo 2π".

Malheureusement, la structure métrique du carré recollé n'est pas représentative de l'idée sous-jacente de "bouée" que nous avons tous en tête : avec un parallélépipède, impossible de rendre le "petit équateur" plus court que le grand – voir Figures 4.12 et 4.13. Pour réunir le meilleur des deux mondes, il faudra *déformer* le carré par une métrique appropriée, qui accorde d'autant plus de poids à $\mathrm{d}\theta$ que le point de base (r, θ) est proche de l'axe de révolution du donut. Vous pourrez trouvez des précisions à ce sujet dans les excellentes notes de Robert T. Jantzen, *Geodesics on the Torus and other Surfaces of Revolution Clarified Using Undergraduate Physics Tricks with Bonus : Nonrelativistic and Relativistic Kepler Problems* : `arxiv.org/abs/1212.6206`.

Cette flexibilité de la géométrie Riemannienne, qui permet de décrire des structures géométriques riches sur un espace d'états restreints (ici, le carré qui peut être déformé en à peu près n'importe quelle surface) de manière intrinsèque, fait tout l'intérêt de la théorie. Nous verrons au chapitre 6 le profit que peut en tirer un mathématicien appliqué.

Conclusion, ouverture vers la géométrie combinatoire

Toutefois, au moment d'écrire ces lignes, il me reste un scrupule. En commençant ce chapitre, je vous ai promis des mathématiques d'aujourd'hui... Qui n'ont à vrai dire toujours pas été abordées ! Alors, certes, la géométrie Riemannienne "classique" présentée ici est nettement postérieure à la géométrie d'Euclide. Mais son versant purement analytique – celui qui se consacre aux beaux résultats bien propres sur les espaces homogènes, le disque de Poincaré – est un pur produit des mathématiques de la fin du XIX$^{\mathrm{e}}$ siècle, de l'âge d'or du *calcul analytique*. Il appartient à ce temps maintenant révolu de l'avant-chaos, de l'avant-Gödel où l'on pouvait encore penser que tout était "déterministe, donc calculable, donc développable en séries entières complexes" – en forçant à peine le trait.

Aujourd'hui, de l'eau a coulé sous les ponts. Bien des mathématiciens se sont succédé, et ont épuisé les problèmes classiques, trouvé tout ce qui était pertinent et accessible au calcul explicite. Surtout, avec les bouleversements de la physique et de l'informatique sont arrivés de nouveaux problèmes qui venaient de la mécanique quantique ou relativiste, de la mécanique des fluides...

Les intérêts des géomètres ont donc considérablement changé. Pour clore cette séance, je voudrais vous présenter une piste, une généralisation féconde de la géométrie de Poincaré qui a connu de grands développements ces trente dernières années.

Tout est parti d'un constat simple : dans le plan hyperbolique – i.e. le disque de Poincaré, l'hémisphère ou le demi-plan, qui sont équivalents les uns aux autres –, les triangles sont uniformément fins.

Proposition 4.2 (Les triangles du plan hyperbolique sont uniformément fins)**.** *Soit a, b, c trois points du plan hyperbolique. On peut tracer les trois segments hyperboliques $s_{a\to b}$, $s_{b\to c}$ et $s_{c\to a}$ – que l'on notera plus simplement $[a, b]$, $[b, c]$ et $[c, a]$ – , pour obtenir le triangle (abc), exact analogue des triangles du plan euclidien. Alors, et c'est un fait remarquable :*

$$\forall\ x \in [a, b],\ \exists\ y \in [a, c] \cup [c, b],\ \mathrm{d}_H(x, y) \leqslant \ln(1 + \sqrt{2}) = 0.88... \tag{4.75}$$

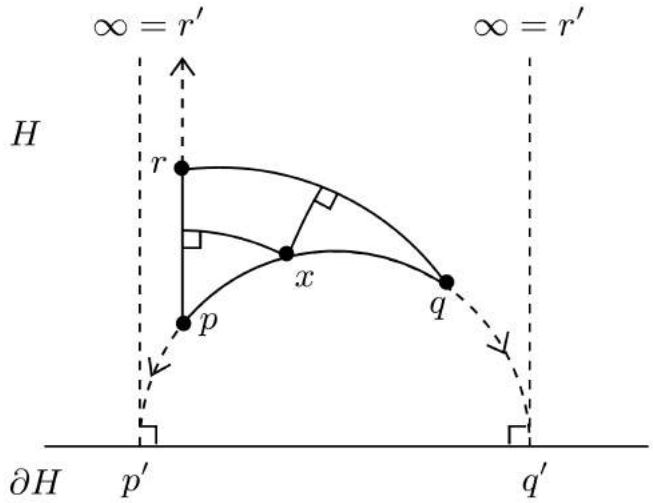


FIGURE 4.14 – Finesse des triangles dans le plan hyperbolique, comme expliqué Proposition 4.2. Prouver l'exactitude de la borne $\ln(1+\sqrt{2})$ demande un peu de travail, mais cette figure permet de comprendre intuitivement pourquoi envoyer les sommets à l'infini ne permet pas de faire croître indéfiniment l'épaisseur des triangles géodésiques.
Un exemple de triangle limite, avec p, q et r à l'infini est ici représenté en pointillé. Notons pour commencer que, puisque la température est proportionnelle à l'altitude sur le demi-plan, la distance entre les deux branches verticales s'amenuise (en tendant vers 0) à mesure que l'on remonte vers $r' = \infty$. Le même pincement des bords est vrai au voisinage des points p' et q'. Comme les dilatations partant du bord sont des isométries, il n'est pas difficile de voir que le point de $s_{p\to q}$ le plus éloigné des deux autres segment se trouve au milieu de l'arc de cercle... à distance finie (et très raisonnable) des deux bords verticaux. Envoyer les points à l'infini dans des directions opposées n'a donc pas permis d'épaissir le triangle : c'est une différence remarquable entre le plan hyperbolique et le plan euclidien.
Dessin tiré de l'article de J. W. Cannon et al., *Hyperbolic Geometry.*

Autrement dit, le plus court chemin pour aller de a à b reste à une distance uniformément bornée par 0.9 *du chemin* $[a, c] + [c, b]$.

Le résultat ci-dessus se comprend par exemple dans le modèle du disque de Poincaré : étant donnés trois points a, b et c "typiques", on voit que les géodésiques auront toujours intérêt à repiquer vers le centre "chaud" du disque. Impossible donc de trouver des côtés bien écartés dans un triangle hyperbolique.

Il s'agit, évidemment, d'une *immense* fracture entre le monde euclidien et le monde hyperbolique. Là où, dans le plan euclidien, il suffit de *dilater* pour changer d'échelle et augmenter le diamètre d'un triangle géodésique, un tel argument n'est plus valide dans le disque de Poincaré. Nous avons d'ailleurs vu que certaines dilatations du demi-plan, loin de les dilater, *conservaient* les distances !

Le disque de Poincaré est donc plein de surprises. Heureusement, les intuitions à son sujet ne manquent pas : je vous conseille à ce propos l'excellent article de vulgarisation d'Étienne Ghys, *Poincaré et son disque.* Au delà des résultats *analytiques*, les mathématiciens vont peu à peu trouver des analogues *discrets* au plan hyperbolique, qui imitent sa structure métrique à la manière d'un quadrillage sur le plan euclidien.

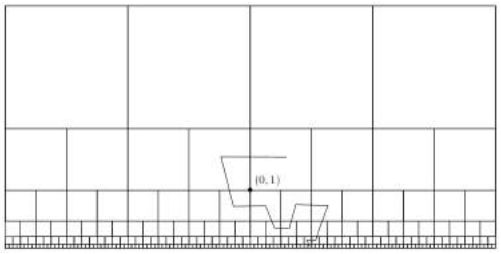
Le sixième modèle

Celui qu'on appellera ici le "sixième modèle" (après I, J, H, et les deux modèles plus anecdotiques du disque de Klein K et de l'hyperboloïde L) est construit par *pavage* du demi-plan, comme à la Figure 4.15. Un nœud du graphe, un carré, sera donc relié à exactement cinq voisins : un en haut, un à gauche, un à droite et... deux en bas ! Les distances sont alors calculées exactement comme on le ferait sur un quadrillage classique, modulo cet *excès de masse* caractéristique de la géométrie hyperbolique.

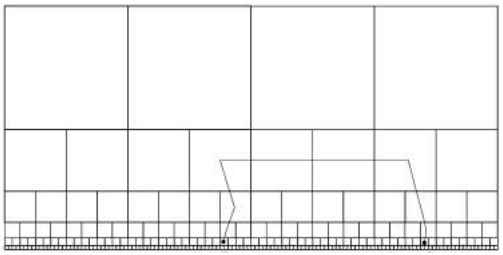
S'il n'est pas parfait, ce modèle approché reflète deux propriétés bien réelles de la géométrie hyperbolique :

— Son homogénéité (qui n'est qu'approchée ici, avec une direction haut-bas privilégiée) : aucun point n'est "au centre" de l'espace, tous sont équivalents au sens de la distance du graphe.

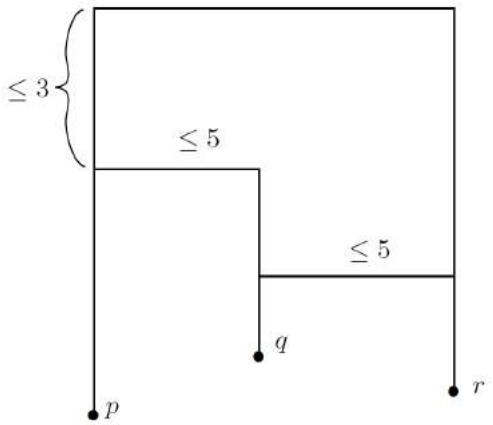
— La finesse de ses triangles, illustrée Figure 4.15 : étant donnés trois points a, b, c du graphe, il n'est pas très difficile de comprendre que les géodésiques ont intérêt à "remonter" à des altitudes similaires, pour profiter des "grands carrés".



(a) Un chemin quelconque dans le sixième modèle.



(b) Géodésique reliant deux points du graphe "proches" du bord.



(c) Plutôt que de couper à travers une forêt de petits carrés, il est souvent préférable de prendre un peu de hauteur, jusqu'à avoir la vue dégagée. Les géodésiques du sixième modèle ne passent donc jamais plus de cinq étapes à se translater sur la gauche ou la droite, préférant à cela la grimpette.

FIGURE 4.15 – Dans le sixième modèle, un triangle géodésique (pqr) ne peut être d'épaisseur supérieure à 8.
Dessins tirés de l'article de J. W. Cannon et al., *Hyperbolic Geometry*.

L'hyperbolicité au sens de Gromov

Ce dernier point de vue, combinatoire, fait ressortir de manière forte la propriété 4.2 sur la finesse des triangles : il s'agit d'une notion qui "passe au discret", et qui ne dépend donc pas de la structure analytique du plan hyperbolique. Malgré son apparente simplicité, elle capture donc l'essentiel des propriétés *métriques* du disque de Poincaré (ou *plan hyperbolique*) dans une proposition qui dépend uniquement des *distances* entre points, et plus des longueurs de chemins.

Une idée extrêmement importante due à Mikhaïl Gromov a été, dans le courant des années 80, de prendre cette propriété pour *définition* de l'hyperbolicité :

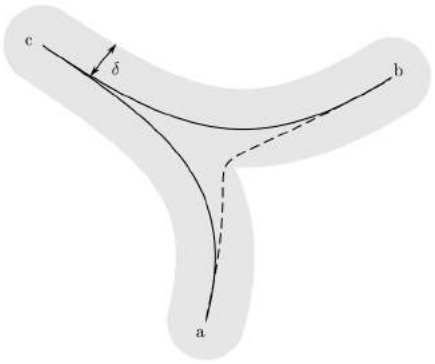


FIGURE 4.16 – Un espace métrique est δ-hyperbolique au sens de Gromov si tous ses triangles sont δ-fins : pour tous points a, b, c, le segment géodésique $[a, b]$ est inclus dans un δ-voisinage de $[a, c] \cup [c, b]$.

Ainsi, le disque de Poincaré est $\ln(1 + \sqrt{2})$-hyperbolique (Proposition 4.2), et le "sixième modèle" est 8-hyperbolique, tandis que le plan euclidien ou le quadrillage régulier $\mathbb{Z} \times \mathbb{Z}$ ne le sont évidemment pas, pour aucune valeur de δ.

Surtout, des exemples tout naturels nous viennent de la théorie des graphes : les arbres, ou graphes acycliques, sont des espaces métriques 0-hyperboliques. En effet, entre deux points a et b quelconque d'un arbre passe un unique chemin sans rebroussement : tout triangle géodésique (abc) est donc absolument plat, avec $[a, b] \subset [a, c] \cup [c, b]$. Dans la vision de Gromov – et du point de vue de la seule structure métrique – **un espace hyperbolique n'est donc rien d'autre qu'un "quasi"-arbre** !

Cela n'a l'air de rien, mais c'est exactement le genre d'idées qui fait la différence entre les mathématiques du XIX^{e} siècle et la géométrie contemporaine. Des belles preuves analytiques, du travail sur les projections stéréographiques accessibles au calcul direct, on est passé à l'étude d'espaces a priori beaucoup moins "sympathiques" et pourtant si naturels : pensez aux pavages d'Escher !

Et les applications dans tout ça ? Vous l'avez déjà deviné : la géométrie Riemannienne est le socle sur lequel repose la théorie de la relativité générale d'Einstein. D'un point de vue "purement" mathématique, elle fournit un vocabulaire adapté à l'étude d'espaces non-plats – aujourd'hui, on verrait par exemple ces espaces "modèles" comme les représentant les plus confortables (car lisses) de "types" géométriques bien définis. Les généralisations, les directions de recherche sont alors nombreuses : description fine des flots de courbures, de l'évolution des métriques dans des espaces de dimension 4 (motivés à l'origine par des questions cosmologiques), extension à la théorie du contrôle optimal (guidage de véhicule, mise en orbite de satellites...).

Dans la fin de ce cours, nous découvrirons un champ d'applications peu médiatisé mais aujourd'hui en plein essor : les extensions à la dimension infinie pour les problèmes de mécanique des fluides, et l'étude des espaces de formes avec applications au traitement d'images médicales.

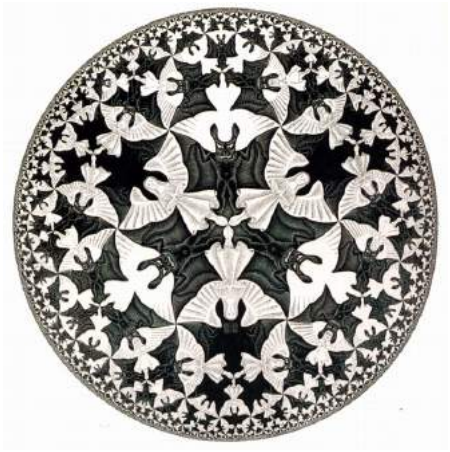


FIGURE 4.17 – Circle Limit IV (Heaven and Hell) de M.C. Escher : que se passerait-il si anges et démons jouaient au téléphone arabe ?

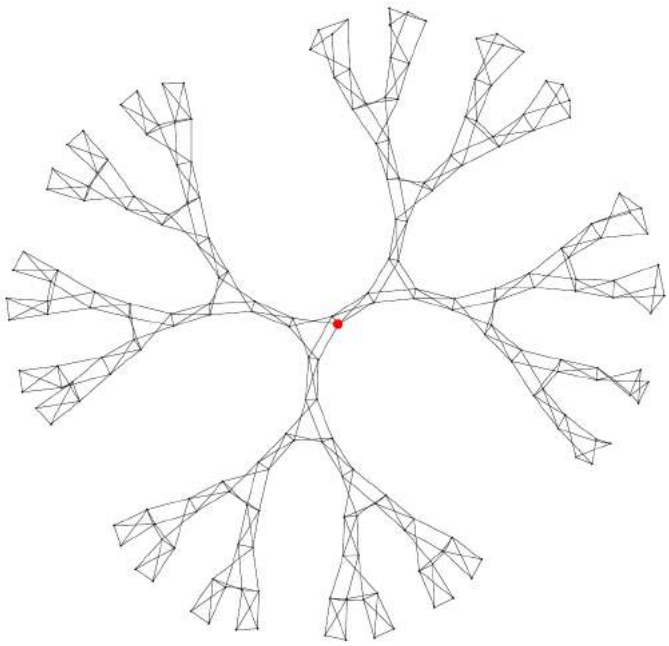


FIGURE 4.18 – Un exemple d'espace métrique hyperbolique non-trivial : le graphe de Cayley (ici tracé au voisinage de l'unité pour un bon système de générateurs) du groupe $SL_2(\mathbb{Z})$, qui agit naturellement sur le demi-plan H par isométries. C'est l'analogue hyperbolique de la grille infinie classique, graphe de Cayley du groupe $\mathbb{Z} \times \mathbb{Z}$ qui agit par isométries sur le plan euclidien. On peut montrer le résultat suivant, parfaite illustration des travaux de l'école de Gromov : "$SL_2(\mathbb{Z})$ étant un groupe *hyperbolique* (i.e. un de ses graphes Cayley a tous ses triangles δ-fins pour une bonne valeur de δ), il est nécessairement *automatique*, c'est-à-dire que sa structure algébrique peut être entièrement encodée dans une collections d'automates finis *sans mémoires*".

Chapitre 5

Un espace de formes étonnant : la sphère des triangles

Séance 6

Le traitement d'images : un domaine qui occupe une place de choix dans ce monde toujours plus numérique. À n'en pas douter, on vous a déjà conté les prouesses du "*Big Data*", des réseaux de neurones et autres algorithmes novateurs qui sont promis à un grand avenir. Mais qui sait, vraiment, de quoi il retourne ?

Avec l'apparition de matériels d'imagerie numérique, trois grandes familles de problèmes stimulent aujourd'hui la recherche :

Les problèmes liés à l'acquisition et au stockage des signaux. Compression, débruitage, super-résolution... Il s'agit de tirer le meilleur parti de notre matériel, pour obtenir à moindre coût des images d'une qualité toujours supérieure.

Les problèmes de classification. Étant donnée une photo arbitraire, est-il possible de deviner ce qu'elle représente ? Est-ce là un avion, une chaise ou un panda ?

Analyse de population. La classification effectuée, comment comprendre, interpréter la variabilité d'une population d'images – de visages par exemple ?

Le modèle universel n'existe pas Avant toute chose, il est important de démystifier nos terminaux numériques. De comprendre que ceux-ci reposent sur des idées, des algorithmes qui n'ont *rien* de magique : une entreprise comme Adobe (éditrice du célèbre logiciel Photoshop) publie ainsi régulièrement des articles scientifiques, emploie des dizaines de chercheurs d'un très haut niveau... qui restent de simples mathématiciens/informaticiens !

En dépit des apparences (s'il y a bien un domaine où l'on reste souvent bluffé par les résultats, c'est celui-là !), les algorithmes les plus sophistiqués reposent donc toujours sur des idées *humainement concevables*. Alors, c'est vrai, de telles idées finissent toujours enrobées de dizaines d'heuristiques, astuces et autres "*tricks*" qui affinent les résultats et permettent de grappiller une poignée de pour-cents sur le produit final... Mais il ne faut pas s'y tromper : le cœur de l'outil est toujours fondé sur la base de quelques équations, quelques idées qui font sens pour les spécialistes.

Qui pourrait alors penser que l'on trouvera un jour "l'équation des images", le "42" des photos numériques ? Les spectaculaires succès de la physique mathématisée ne doivent pas nous griser : si la mécanique céleste peut être résumée en une poignée d'équations, impossible d'espérer la même réussite dans ce domaine-ci. On ne découvrira jamais de "formule" suffisamment simple pour être travaillée, et suffisamment riche pour décrire dans son ensemble la formidable diversité de notre environnement pictural.

Choix d'un modèle adapté Nous avons déjà consacré la section 3.1.3 au problème de la compression d'images, et vu comment l'utilisation d'une transformée de Fourier par blocs était liée un a priori simple :

« Une image naturelle est localement lisse. » (5.1)

En effet, sous réserve que cette hypothèse soit vérifiée, les coefficients associés aux hautes fréquences seront faibles sur chaque bloc de 8x8 pixels et une compression JPEG donnera de bons résultats. Une photo-souvenir "typique" comprenant de grandes plages de dégradés – visages, ciel, ... – on comprend que ce format ait acquis une telle popularité sur les terminaux numériques grand public.

Tout cela est bien beau. Mais il faut garder à l'esprit que si l'hypothèse (5.1) est raisonnable dans le cas de simples "photos", elle est archi-fausse lorsqu'il s'agit de comprimer des images "non-lisses" comme des dessins – voir Figure 5.1 – ou des images très texturées, détaillées comme des scans médicaux. À l'inverse, pour la compression de visages, il est possible vous vous en doutez d'implémenter des a priori de régularité bien plus forts, jusqu'à "faire tenir" une photo d'identité sur un simple QR code comme on l'a vu Figure 3.22 !

À chaque domaine d'application ses a priori, ses contraintes. Confronté à un problème scientifique ou industriel, le mathématicien appliqué aura donc à trouver un cadre, un jeu d'hypothèses pertinentes qui le guident vers des algorithmes aussi efficaces que possible.

Classification par réseaux de neurones Un domaine qui a longtemps résisté aux efforts des spécialistes est celui de la classification d'images naturelles. Le problème était de taille : comment regrouper sous une même étiquette des images qui peuvent différer par l'orientation, l'éclairage, la pose, voire même le style graphique ? Pixel-par-pixel, rien de plus éloigné d'une photo de Concorde qu'un dessin d'avion !

Popularisés sous l'étiquette "Deep Learning" depuis quelques années, les fameux *réseaux de neurones* sont des structures algorithmiques qui encouragent naturellement la création d'invariants aux aléas énumérés plus haut. "Boostés" par l'apparition de puces informatiques adéquates, les fameuses *cartes graphiques* (développées initialement pour l'industrie du jeu vidéo), ces techniques permettent par exemple de décomposer une image en une composante "de texture" et un vecteur "de classe" (indépendant de la texture, de l'éclairage, des petites déformations...), ce qui ouvre la voie à des algorithmes de classification efficaces ou à des applications plus exotiques comme le "transfert de style" présenté sur le site `deepart.io`.

Comment étudier une population d'images semblables ? En résumé, les réseaux de neurones classiques sont des structures taillées pour créer de l'invariance, pour *quotienter* une certaine variabilité des images et n'en conserver qu'une classification utile. Si l'intérêt est évident dans de nombreux cas de figure, il ne sauraient donc être les seuls outils utilisés pour l'analyse de population, où il s'agit justement d'étudier un seul type d'images, une seule *classe*.

Je vous propose aujourd'hui de découvrir les concepts développés depuis les années 40 par les mathématiciens pour attaquer ce problème, sous sa forme la plus générique – dans des cas très spécifiques comme l'analyse de visages, une théorie ad hoc donnera toujours les meilleurs résultats. De manière étonnante, nous nous placerons dans la continuité de la séance précédente et nous retrouverons en fait à étudier "la variété Riemannienne des formes" !

(a) À la transition entre la montgolfière et le ciel bleu, l'élimination des hautes fréquences produit un phénomène de *ringing*, d'oscillations caractéristiques d'un filtre passe-bas.

(b) Un scan de texte, un dessin sont par essence remplis de discontinuités. Une compression JPEG de telles images donne donc, sans surprise, des résultats désastreux.

FIGURE 5.1 – L'efficacité du format JPEG repose sur une hypothèse de régularité de l'image, et comprime l'image en écrasant les hautes fréquences. Sur des images aux transitions brutales – des dessins ou du texte, par exemple – il est donc déconseillé de l'utiliser.
Image tirées de la page web suivante :
`cscie12.dce.harvard.edu/lecture_notes/2015/20150301/handout.html`.

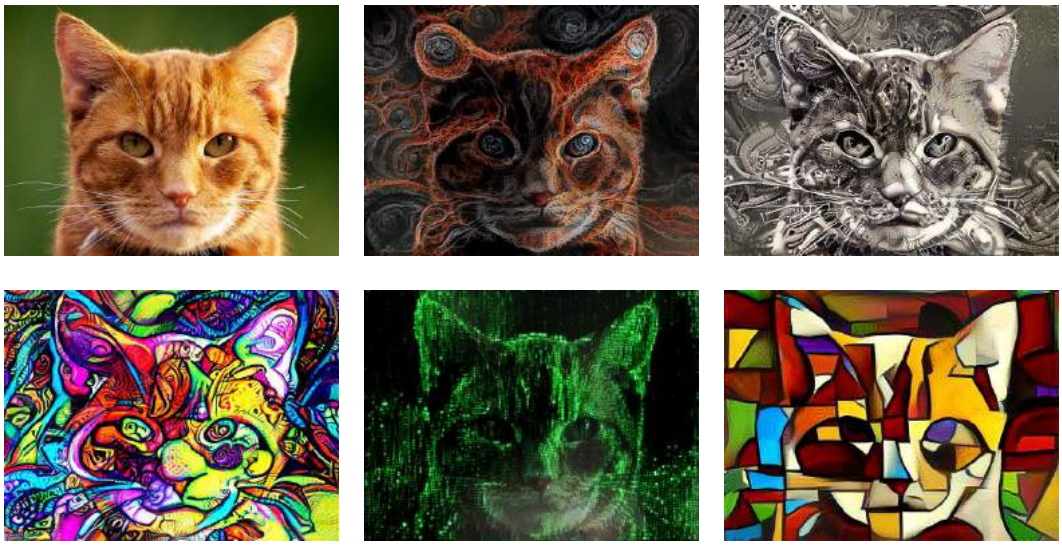


FIGURE 5.2 – Depuis quelques années, on arrive à décomposer une image en une composante de *texture* et un résiduel de *structure*, à l'aide d'algorithmes d'apprentissage utilisant une architecture en cascade plus ou moins inspirée de celle du cortex visuel – d'où le nom de réseaux de neurones. On peut ainsi transférer des "textures" d'images variées sur la structure d'une photo arbitraire, comme le chat présenté en haut à gauche.
Images tirées d'un pre-print paru en 2016, *Exploring the Neural Algorithm of Artistic Style* de Yaroslav Nikulin et Roman Novak.

Étude rudimentaire d'une population de poissons

Commençons par le cas d'une étude écologique menée sur un échantillon de N individus (animaux, feuilles...) photographiés sur le terrain. Par un travail manuel (en voie d'être automatisé aujourd'hui), un expert peut placer sur chaque image des *points de contrôle* bien définis anatomiquement. Bouts de nez, pointes de nageoires, troisièmes nervures... Étiquetés, ces "*landmarks*" permettent de mettre en correspondance les régions analogues d'un individu à un autre. Après un pré-traitement illustré Figure 5.5a, on peut considérer que l'on dispose en fait de population d'une collection $P^1, \dots, P^N$ de nuages de I points de $\mathbb{R}^d$, où N est le nombre d'individus, I le nombre de landmarks par individu et d le nombre de coordonnées par point (typiquement 2, sur des photos).

Action des similitudes Ces nuages de points P^n, on peut les transformer à l'aide de translations, homothéties, rotations : si $(\overline{x}, \overline{y})$ est un vecteur du plan, s est une échelle et θ un angle, on définit l'application

$$S_{\overline{x},\overline{y},s,\theta} : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} s\cos(\theta) \cdot x - s\sin(\theta) \cdot y \ + \ \overline{x} \\ s\sin(\theta) \cdot x + s\cos(\theta) \cdot y \ + \ \overline{y} \end{pmatrix} \tag{5.2}$$

qui est la composée d'une translation de vecteur $(\overline{x}, \overline{y})$, d'une rotation d'angle θ et d'une homothétie de rapport s strictement positif. À vrai dire, puisque j'ai choisi de vous parler aujourd'hui de photos bi-dimensionnelles, on peut se simplifier la vie en adoptant le formalisme complexe. Comme au chapitre 2, on caractérisera nos points (x, y) par des affixes $z = x + iy$. Une similitude du plan sera alors la donnée de deux complexes

$$\tau = \overline{x} + i\overline{y} \qquad \text{et} \qquad \upsilon = se^{i\theta}, \tag{5.3}$$

et on pourra écrire nos similitudes comme des applications

$$S_{\tau,\upsilon} : z \mapsto \upsilon \cdot z + \tau. \tag{5.4}$$

Si $P^n = (P_1^n, \dots, P_I^n)$ est l'une de nos "formes", on définira sans ambages son transformé :

$$S_{\tau,\upsilon}(P^n) = (S_{\tau,\upsilon}(P_1^n), \dots, S_{\tau,\upsilon}(P_I^n)) \,. \tag{5.5}$$

Si P^m et P^n sont deux individus distincts, ne pourrait-on alors tenter de *recaler* l'un sur l'autre à l'aide d'une similitude ? Dans l'esprit des pages précédentes, cela permettrait de décomposer la différence $P^m - P^n$ entre deux nuages en une composante *rigide* $P^m - S_{\tau,\upsilon}(P^m)$, donnée de $(\overline{x}, \overline{y})$, s et θ, et un *résiduel* irréductible $S_{\tau,\upsilon}(P^m) - P^n$.

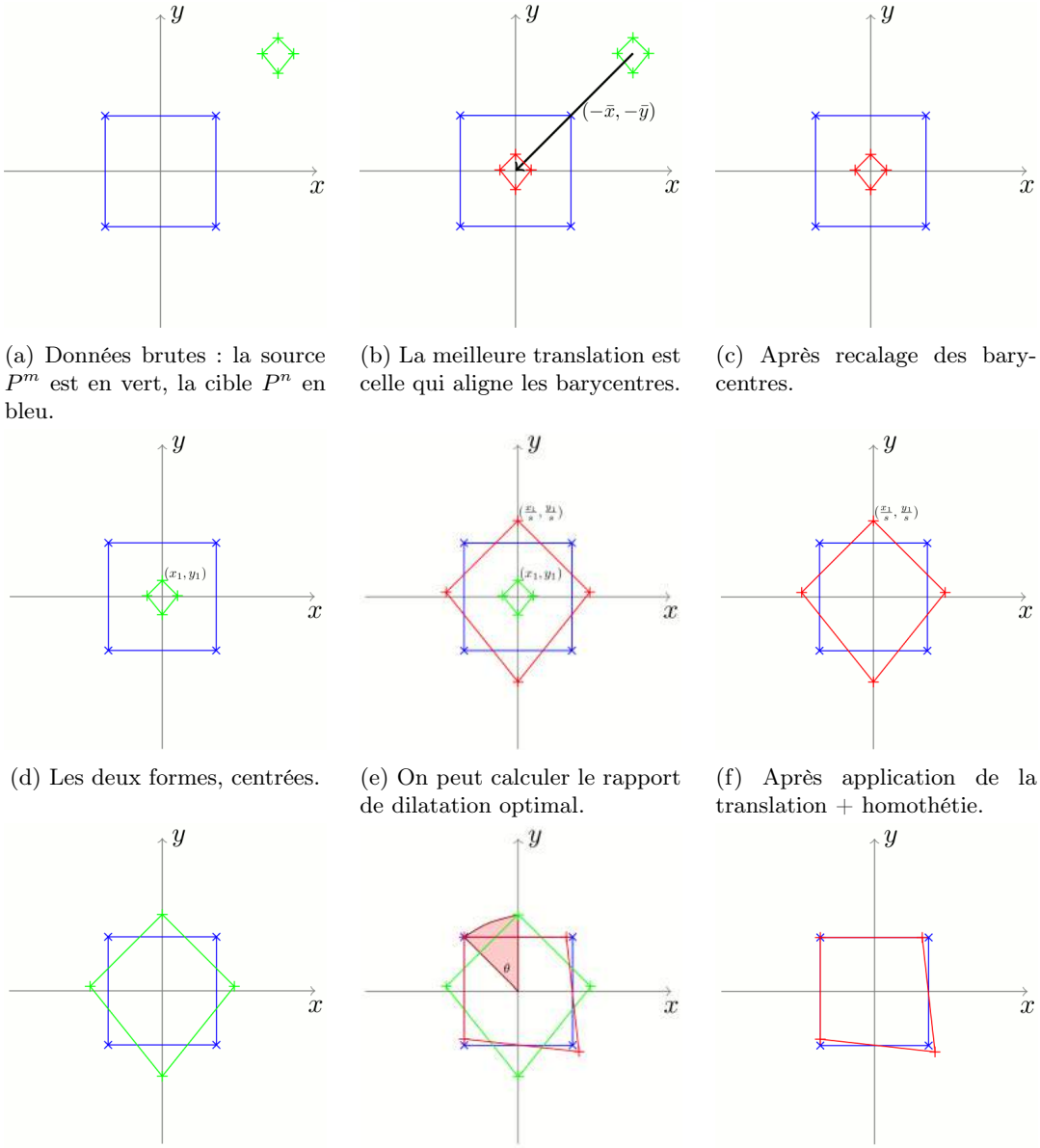
Illustré Figure 5.3, le procédé le plus simple est de trouver le choix de (τ, υ) qui minimise

$$\|S_{\tau,\upsilon}(P^m) - P^n\|_2^2 \ = \ \sum_{i=1}^{I} |S_{\tau,\upsilon}(P_i^m) - P_i^n|^2 \tag{5.6}$$

$$= \ |\upsilon \cdot P_1^m + \tau - P_1^n|^2 + \cdots + |\upsilon \cdot P_I^m + \tau - P_I^n|^2 \,. \tag{5.7}$$

Cette méthode a été popularisée sous le nom d'analyse *Procustéenne* en référence à Procuste, brigand apparaissant dans la légende de Thésée qui, d'après Diodore de Sicile

« contraignait les voyageurs à se jeter sur un lit ; il leur coupait les membres trop grands et qui dépassaient du lit ; et étirait les pieds de ceux qui étaient trop petits. »



(a) Données brutes : la source P^m est en vert, la cible P^n en bleu.

(b) La meilleure translation est celle qui aligne les barycentres.

(c) Après recalage des barycentres.

(d) Les deux formes, centrées.

(e) On peut calculer le rapport de dilatation optimal.

(f) Après application de la translation + homothétie.

(g) Les deux formes sont centrées et normalisées : reste à recaler les orientations.

(h) La meilleure rotation s'obtient elle aussi par un calcul analytique simple.

(i) Après recalage par une similitude : le résidu peut être analysé indépendamment.

FIGURE 5.3 – L'analyse Procustéenne, ou la recherche pas-à-pas de la meilleure similitude pour *recaler* deux formes – ici, un quadrilatère vert sur le carré bleu. Les paramètres optimaux pour la similitude sont obtenus par un calcul élémentaire, détaillé dans le cas des triangles aux équations (5.21-5.50). Après recalage (en bas à droite), les deux quadrilatères diffèrent uniquement par une fine différence de *forme* (au sens usuel) : la différence qui existe entre un cerf-volant et un carré, indépendante de tout problème d'échelle ou d'orientation.
Images tirées de Wikipédia, par Linschn.

Un outil de première analyse En Figure 5.4, on présente un exemple de recalage procustéen entre deux ailes de libellules. Par une procédure de minimisation simple, on réussit à trouver les paramètres d'échelle, d'orientation et de translation qui décrivent au mieux le passage entre deux formes anatomiques quelconques. Appliqué à une population de silhouettes, ceci permet par exemple de tracer des courbes de croissances.

Un outil de pré-traitement indispensable Mais si l'analyse rigide est toujours étudiée aujourd'hui, c'est avant tout parce qu'elle permet de travailler sur des résiduels *propres*. En Figure 5.5, j'ai reproduit une collection d'images tirées d'un article d'écologie "de routine", *A morphometric approach for the analysis of body shape in bluefin tuna : preliminary results*, d'Addis, Melis, Cannas, et al., paru en 2009. Ayant capturé et photographié deux cohortes d'environ 60 thons rouges méditerranéens en 2008 et 2009, les chercheurs ont analysé les deux populations.

Pour résumer, après positionnement de points de contrôles anatomiques, l'analyse de forme se fait en deux temps. Une analyse procustéenne est d'abord conduite pour éliminer les paramètres de cadrage (translation, rotation) et de taille (échelle), jugés peu significatifs : ceux-ci sont directement influencés par les prises de vues expérimentales. Les données *recalées par des similitudes* sur un modèle de référence sont alors étudiées au travers de déformations en "plaques minces" – un modèle simple de déformations élastiques, mis à disposition de tous dans le manuel *Morphometric tools for landmark data : geometry and biology* de Bookstein, paru en 1991. In fine, l'analyse semble dégager une variation significative de morphologie entre les deux populations, correspondant à une certaine cambrure des poissons pêchés en 2009 : les auteurs de l'étude suggèrent de la lier au cycle reproductif de l'espèce.

Cet article permet d'illustrer le besoin d'outils fiables et riches pour l'analyse automatique de formes. Si nous étudierons les outils fins (semblables à l'analyse en plaques minces) au chapitre suivant, le recalage par similitudes a bien joué ici un rôle clé : en *quotientant* les déformations rigides, ils nous permet de travailler sur des *classes d'équivalences* de poissons à similitude près, "types" de formes abstraits décorrélés des considérations de prise de vue ou de gabarit.

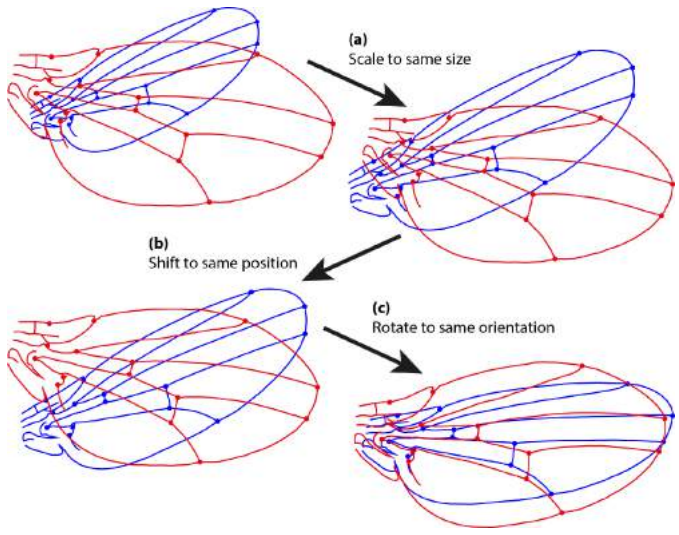
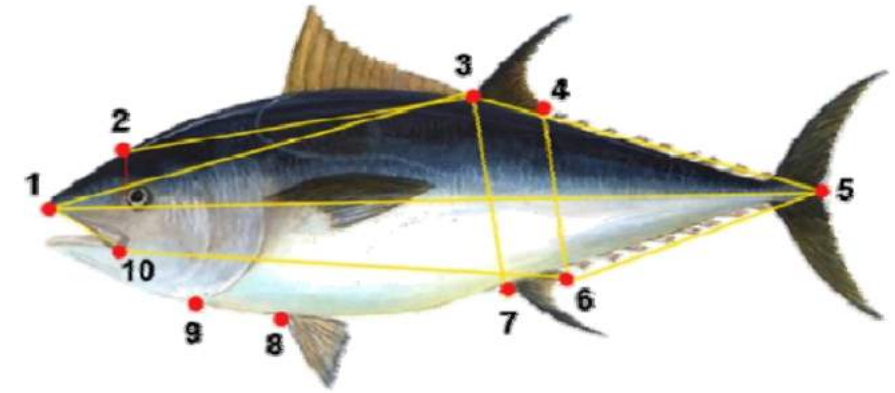
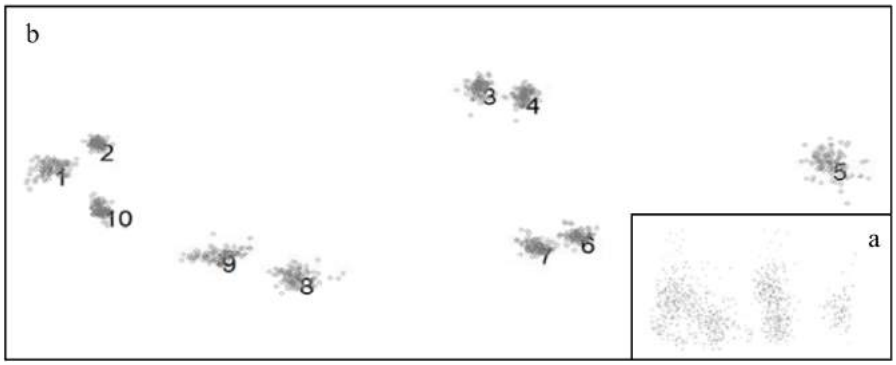


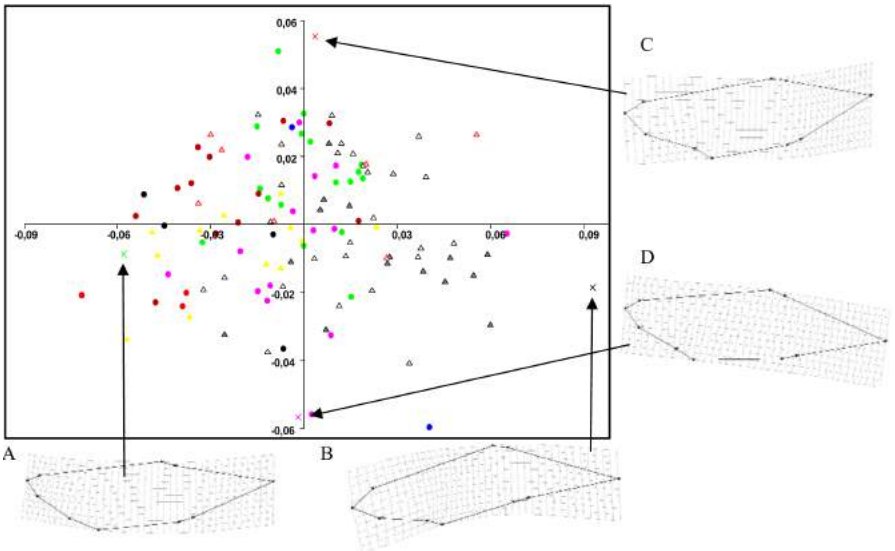
FIGURE 5.4 – Recalage de l'aile bleue sur l'aile rouge par une analyse procrustéenne : celle-ci décompose la variation inter-sujets en une similitude – position, orientation, taille – et un résiduel. Selon les cas, il sera pertinent d'étudier l'une de ces deux composantes et d'oublier l'autre. Image tirée de Wikipédia.



(a) Landmarks anatomiques sur un thon.



(b) Pertinence de l'analyse procrustéenne : le nuage de données brutes est représenté en vignette (a), en bas à droite. Quotienter par les similitudes permet de recaler tous les nuages de 10 points autour d'une même forme de référence, en faisant abstraction de la position (abscisse et ordonnée), de l'orientation ou de l'échelle. On passe donc de 20 à 16 degrés de liberté indépendants. L'analyse de forme se fera sur ces données "propres" (b), robustes vis à vis des problèmes de cadrage au moment de l'acquisition.



(c) Après une analyse en "plaques minces", les deux principaux modes de variation de nos formes de thons "à similitude près" sont isolés, et servent à repérer les poissons dans le plan ci-dessus : le premier mode correspond ainsi à la direction (AB), et le second mode à la direction (CD). Les disques correspondent aux animaux pêchés en 2008 et les triangles au millésime 2009 : une inhomogénéité manifeste est révélée, avec la sur-représentation des poissons de 2009 dans le quadrant inférieur droit.

FIGURE 5.5 – Analyse morphométrique rudimentaire d'une population de thons.

Menhirs, Cornouailles et sphère des triangles

L'analyse procustéenne n'est pas qu'une routine de pré-traitement élémentaire. De manière surprenante, elle a motivé un bel effort de recherche dans le courant des années 80 avec en point d'orgue l'étonnant résultat de David G. Kendall : la compréhension définitive de l'espace des formes de triangles, isométrique à une sphère (!) de rayon 1/2.

Illustrée en Figure 5.6, la première motivation de l'étude de l'espace des triangles à similitude près est la suivante :

« Étant donnée une collection arbitraire de points du plan, à partir de quels critères peut-on affirmer qu'elle présente "trop" d'alignements, indices d'une structuration sous-jacente ? »

L'*alignement* étant une question indépendante de l'échelle ou de l'orientation des données, il est naturel de travailler sur cette question "à similitude près". Comme elle peut s'exprimer en terme de triplets de points, elle nous incite à concentrer nos efforts sur le cas des triangles $P^n = (P_1^n, P_2^n, P_3^n)$. À partir d'une collection de points, on peut construire l'ensemble de tous les triangles y prenant leurs sommets, et donc se ramener à une question bien posée :

« Quelle est la distribution "standard" des formes de triangles – à similitude près – reliant des points tirés au hasard selon une loi de probabilité simple ?
Plus spécifiquement, si des points sont (par exemple) tirés de manière indépendante et uniforme dans le disque unité, puis reliés trois à trois, quelle sera la proportion observée de triangles "presque plats" avec un angle inférieur à 1° ? »

(a) À l'extrémité ouest des îles britanniques, la pointe des Cornouailles recèlerait-elle un secret ?
Carte OpenStreetMap.

(b) Sur cette péninsule de 10km sur 15km, on observe une concentration élevée de monolithes.

FIGURE 5.6 – Motivation initiale de David Kendall : les monolithes de Land's End ont-ils été placés au hasard, ou peut-on voir dans la distribution ci-dessus les traces d'alignements significatifs ? Chaque site est ici représenté par un nombre, celui des "triangles presque plats" dont il fait partie. On reformule alors notre question de manière statistique :
« Sur la carte (b), y a-t-il un nombre anormalement élevé de *triades* de mégalithes alignés ? »
Dans un article d'une grande importance historique, *Simulating the Ley Hunter* paru en 1980, Simon Broadbent laisse à penser que non – ce qui est bien entendu sujet à de sérieuses controverses archéologiques.

Un système de coordonnées adaptées

On va présenter le cadre théorique permettant de répondre de manière élégante à de telles questions. Avant tout, il importe de *représenter* nos triangles sous une forme adaptée aux calculs de **distances** et de **similitudes**.

J'adopterai ici les notations de l'article de référence de David Kendall, *Shape Manifolds, Procustean Metrics and Complex Projective Spaces* (1984), sur lequel est basé la présente discussion. Pour une exposition plus élémentaire, on pourra aussi consulter *Exact Distributions for Shapes of Random Triangles in Convex Sets* du même auteur, publié en 1985 à destination d'un public non spécialisé.

Paramétrisation naïve À première vue, un triangle ABC n'est rien d'autre que la donnée de trois affixes complexes associées aux sommets : $z_1^\star$, $z_2^\star$ et $z_3^\star$. Conformément à l'équation (5.5), pour tout "vecteur" τ et rapport υ, on peut écrire

$$S_{\tau,\upsilon}(ABC) \;=\; S_{\tau,\upsilon}(z_1^\star, z_2^\star, z_3^\star) \;=\; (\; \upsilon \cdot z_1^\star + \tau,\; \upsilon \cdot z_2^\star + \tau,\; \upsilon \cdot z_3^\star + \tau \;). \tag{5.8}$$

De plus, pour $(z_1^\star, z_2^\star, z_3^\star)$ et $(w_1^\star, w_2^\star, w_3^\star)$ deux triangles quelconques, le critère minimisé par l'analyse procustéenne s'écrit simplement :

$$\| z^\star - w^\star \|_2^2 \;=\; |z_1^\star - w_1^\star|^2 + |z_2^\star - w_2^\star|^2 + |z_3^\star - w_3^\star|^2. \tag{5.9}$$

Coordonnées barycentriques Grâce à la multiplication complexe, la représentation naïve des triangles par un triplet d'affixes n'est donc pas un mauvais choix. Mais il est possible d'aller plus loin en considérant le nouveau triplet de coefficients

$$\begin{pmatrix} z_0 & z_1 & z_2 \end{pmatrix} \;=\; \begin{pmatrix} \frac{1}{\sqrt{3}}(z_1^\star + z_2^\star + z_3^\star) & \frac{1}{\sqrt{2}}(z_2^\star - z_1^\star) & \frac{1}{\sqrt{6}}(2z_3^\star - z_1^\star - z_2^\star) \end{pmatrix} \tag{5.10}$$

$$\;=\; \begin{pmatrix} z_1^\star & z_2^\star & z_3^\star \end{pmatrix} \cdot \begin{pmatrix} 1/\sqrt{3} & -1/\sqrt{2} & -1/\sqrt{6} \\ 1/\sqrt{3} & +1/\sqrt{2} & -1/\sqrt{6} \\ 1/\sqrt{3} & 0 & +2/\sqrt{6} \end{pmatrix}. \tag{5.11}$$

Illustré Figure 5.7, ce changement linéaire de coordonnées permet d'écrire les similitudes de manière plus compacte

$$S_{\tau,\upsilon}(ABC) \;=\; S_{\tau,\upsilon}(z_0, z_1, z_2) \;=\; (\; \upsilon \cdot z_0 + \sqrt{3}\,\tau,\; \upsilon \cdot z_1,\; \upsilon \cdot z_2 \;), \tag{5.12}$$

les translations n'ayant d'influence que sur le seul barycentre. Dans le même temps, grâce aux propriétés d'orthogonalité de la matrice de changement de repère, on a toujours une expression simple de la distance quadratique entre deux triangles du plan :

$$\| z^\star - w^\star \|_2^2 \;=\; \|z - w\|_2^2 \;=\; |z_0 - w_0|^2 + |z_1 - w_1|^2 + |z_2 - w_2|^2. \tag{5.13}$$

Relation d'équivalence Si (z_0, z_1, z_2) et (w_0, w_1, w_2) sont deux triangles donnés en coordonnées "barycentriques", on peut maintenant déterminer s'ils sont ou non similaires l'un à l'autre :

$$\exists\, \tau, \upsilon \in \mathbb{C}, S_{\tau,\upsilon}(z) = w \iff \exists\, \tau, \upsilon \in \mathbb{C}, \left\{ \begin{array}{lcccl} \upsilon z_0 & + & \sqrt{3}\,\tau & = & w_0 \\ \upsilon z_1 & & & = & w_1 \\ \upsilon z_2 & & & = & w_2 \end{array} \right. \tag{5.14}$$

$$\iff z_1 w_2 = z_2 w_1. \tag{5.15}$$

Autrement dit, si notre triangle n'est pas dégénéré ($A = B = C$, cas que l'on exclut de notre analyse), on peut caractériser sa forme "à similitude près" par le ratio

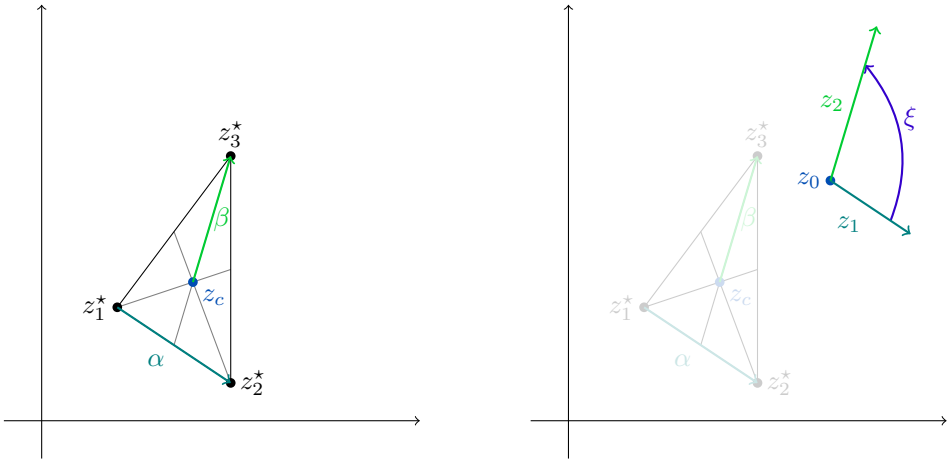
$$\xi \;=\; \frac{z_2}{z_1} \;=\; \frac{\frac{1}{\sqrt{6}}(2z_3^\star - z_1^\star - z_2^\star)}{\frac{1}{\sqrt{2}}(z_2^\star - z_1^\star)} \in \mathbb{C} \cup \{\infty\}. \tag{5.16}$$

Deux triangles non dégénérés (z) et (w) seront semblables si et seulement si leurs "ratios" respectifs $\xi = z_2/z_1$ et $\xi' = w_2/w_1$ sont égaux. On a réussi à décrire la forme d'un triangle par un unique nombre complexe "projectif" (i.e. qui peut être infini), avec *deux* paramètres réels (module et argument). À vrai dire, comme illustré Figures 5.8 et 5.9, considérer le ratio ξ revient ni plus ni moins à utiliser une similitude pour recaler $(A, B) = (z_1^\star, z_2^\star)$ sur $(-1, 1)$, puis à identifier la forme du triangle avec la coordonnée libre restante, issue de $z_3^\star$.

Premiers exemples Certaines valeurs remarquables de ξ correspondent à des propriétés simples du triangle ABC ; on trouve :

— $\xi = \infty \quad \Leftrightarrow \; A = B.$
— $\xi = 0 \quad \Leftrightarrow \; C$ est le milieu de $[AB]$.
— $\xi \in \mathbb{R} \cup \{\infty\} \quad \Leftrightarrow \; ABC$ est plat.
— $\xi \in i\mathbb{R} \cup \{\infty\} \quad \Leftrightarrow \; ABC$ est isocèle en C.
— $\xi \in \{-i, +i\} \quad \Leftrightarrow \; ABC$ est équilatéral.

Malheureusement, manipuler ξ manque d'élégance : pourquoi mettre à part le triangle avec $A = B$, correspondant à l'infini ? L'espace des triangles devrait refléter une certaine invariance par permutation des sommets, ce qui n'est clairement pas le cas ici. Pour trouver une représentation satisfaisante de l'espace des formes de triangles, il faut donc travailler encore un peu.



(a) Des affixes aux coordonnées barycentriques. (b) Des coordonnées barycentriques au ratio projectif.

FIGURE 5.7 – On peut caractériser un triangle $(z_1^\star, z_2^\star, z_3^\star)$ par son barycentre $z_c = \frac{1}{3}(z_1^\star + z_2^\star + z_3^\star)$ et deux vecteurs directionnels $\alpha = z_2^\star - z_1^\star$ et $\beta = \frac{1}{3}(2z_3^\star - z_1^\star - z_2^\star)$. En normalisant ceux-ci pour obtenir un jeu de coordonnées $(z_0, z_1, z_2) = (\sqrt{3}\, z_c, \alpha/\sqrt{2}, \beta\sqrt{3}/\sqrt{2})$, on dispose d'une représentation adaptée aux similitudes *et* aux calculs de distances – équations (5.12-5.13). Finalement, le ratio $\xi = z_2/z_1$ encode complètement la forme du triangle aux similitudes près.

La sphère des triangles Souvenons-nous : dans le film *Dimensions* repris Figure 2.5b, nous avions appris qu'une "droite projective complexe" n'est rien d'autre qu'une sphère. Par la projection stéréographique, il est donc possible d'identifier un ratio complexe $\xi \in \mathbb{C} \cup \{\infty\}$ avec un point de la boule unité, caractérisé par deux angles en coordonnées sphériques θ et φ. En s'inspirant de cette idée, David Kendall propose d'associer un couple d'angles à chaque triangle $(z^\star) \sim (z) \sim \xi$ par la formule :

$$\theta \;=\; 2\arctan(|\xi|) \qquad \text{et} \qquad \varphi \;=\; \arg(\xi), \tag{5.17}$$

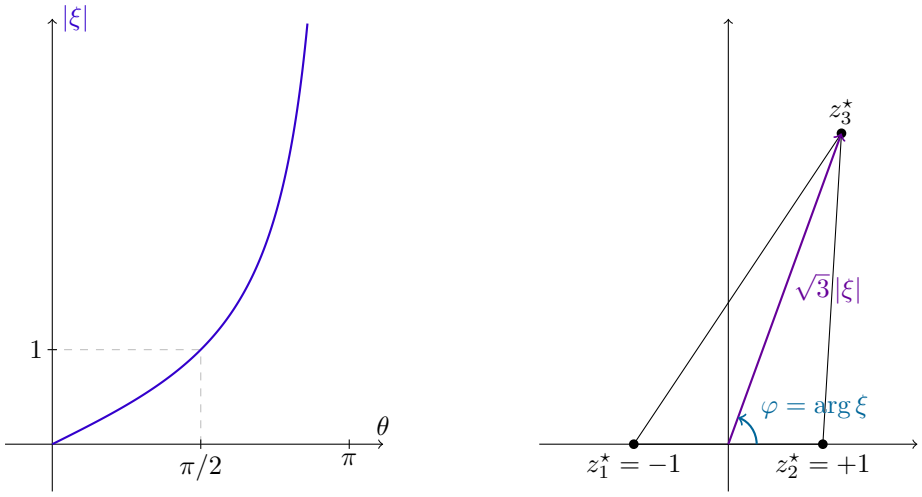
$$\text{de sorte que} \qquad \xi = \tan(\theta/2)\, e^{i\varphi}. \tag{5.18}$$

Chaque classe de triangle à similitude près est alors associée à un point de la sphère, que l'on choisit de rayon $1/2$ pour une raison qui deviendra claire au Théorème 5.1 :

$$f(\xi) \;=\; \frac{1}{2}(\cos\theta, \sin\theta\cos\varphi, \sin\theta\sin\varphi) \tag{5.19}$$

en coordonnées (X, Y, Z). Le résultat est illustré Figure 5.10, avec un représentant de classe pour chaque point du globe. Suivant la page précédente, on trouve :

— Le segment $[A = B, C]$ au point $(-1/2, 0, 0)$.
— Le triangle plat où C est au milieu de $[AB]$ au point $(+1/2, 0, 0)$.
— Les triangles plats sur l'équateur $Z = 0$.
— Les triangles isocèles en C sur le méridien $Y = 0$.
— Les deux triangles équilatéraux direct (Nord) et indirect (Sud) aux pôles $(0, 0, \pm 1/2)$.



(a) $|\xi|$ en fonction de θ.

(b) Construction d'un triangle de coordonnées sphériques (θ, φ).

FIGURE 5.8 – Étant donné un point de la sphère de coordonnées sphériques $(\theta, \varphi) \in [0, \pi] \times [0, 2\pi[$, comment construire un triangle lui correspondant ? Il suffit de considérer le triplet $(z_1^\star, z_2^\star, z_3^\star) = (-1, 1, \sqrt{3}\,\tan(\theta/2)\, e^{i\varphi})$.

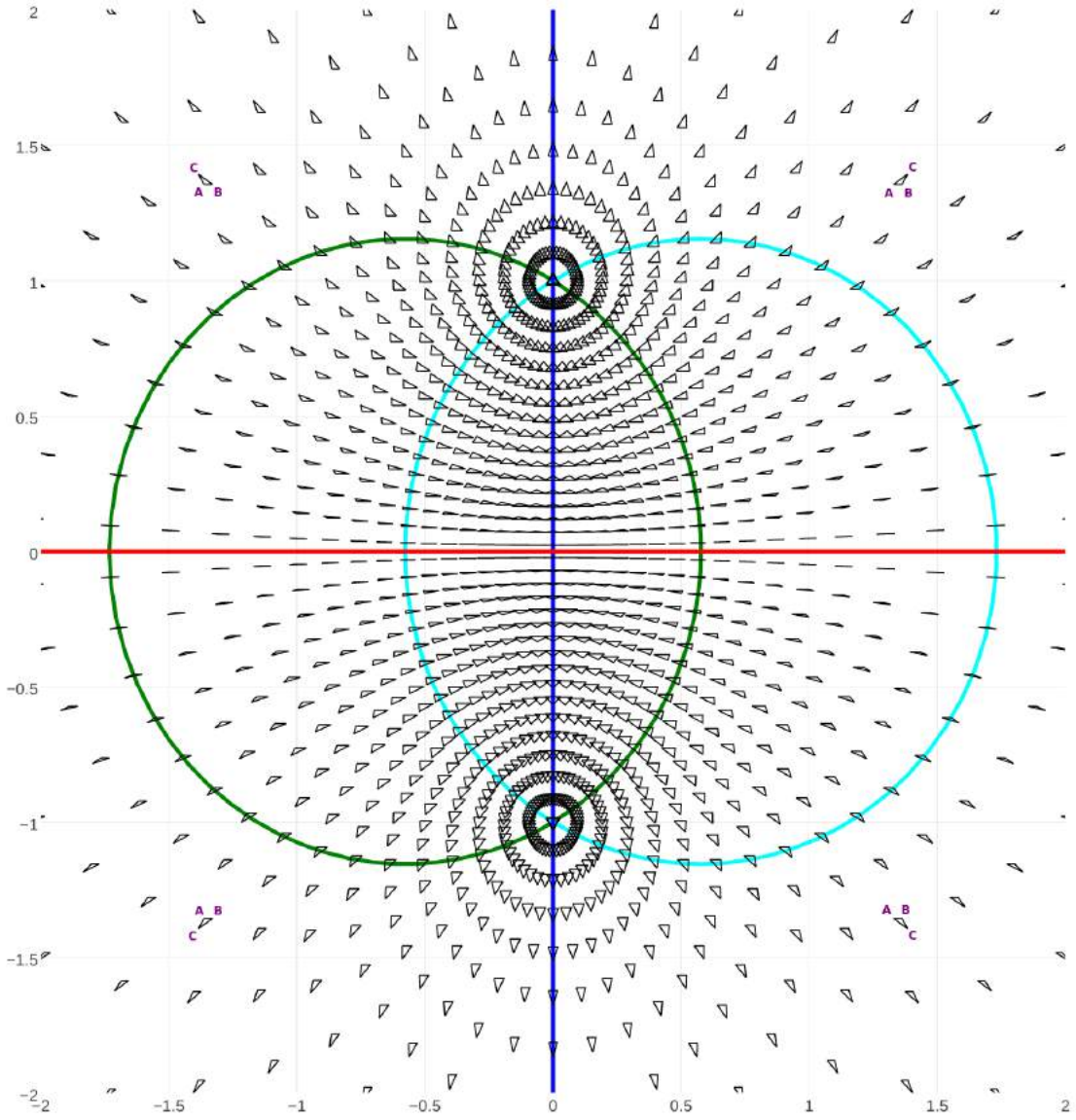
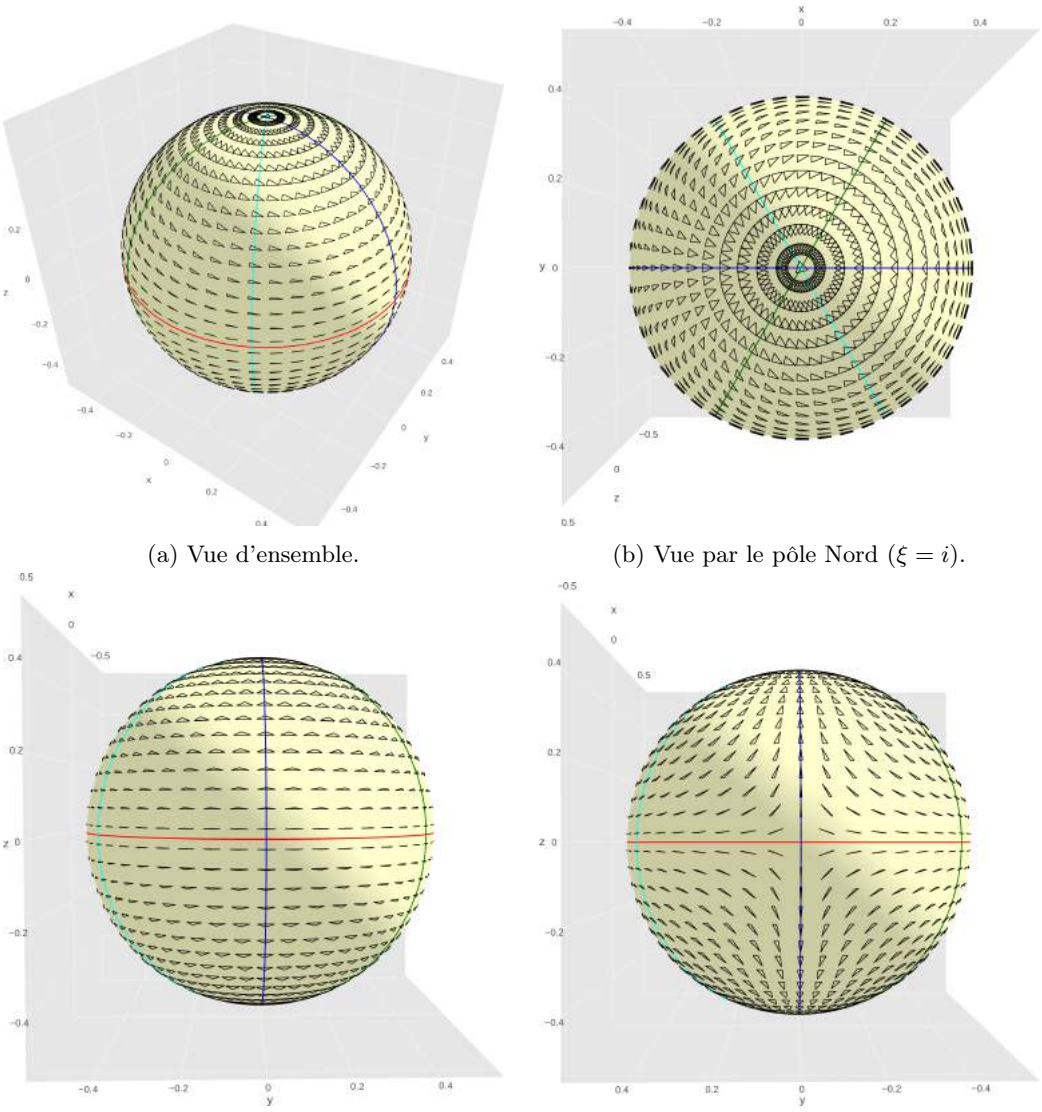


FIGURE 5.9 – L'espace des triangles à similitude près, vu dans le plan des ratios $\xi \in \mathbb{C} \cup \{\infty\}$. Dessiné en rouge, l'axe des abscisses correspond aux triangles plats. Les triangles isocèles se répartissent selon trois courbes, qui dépendent du sommet privilégié : en bleu, l'axe des ordonnées correspondant aux triangles isocèles en C ; en vert la courbe des triangles isocèles en A, en cyan celle des triangles isocèles en B. Le point correspondant au ratio $\xi = 0$ est ici représenté à l'origine du repère.



(a) Vue d'ensemble.

(b) Vue par le pôle Nord ($\xi = i$).

(c) Vue du golfe de Guinée.

(d) Vue de Tahiti.

FIGURE 5.10 – L'espace des triangles à similitude près, isométrique à la sphère de rayon 1/2. Homogène, cette représentation permet de rendre compte des symétries intrinsèques de l'espace des triangles liées au ré-étiquetage des points et aux réflexions. Ici, chaque "forme" au sens des rapports de longueurs se retrouve dans les 12 quartiers découpés par l'équateur des triangles plats et les méridiens des triangles isocèles.

Sphère des triangles et distance procustéenne

La sphère des triangles ne pourrait être qu'un moyen astucieux de rendre compte de la *topologie* de l'ensemble des triangles considérés à similitude près. Un moyen commode de se souvenir que pour passer d'un triangle équilatéral direct à un triangle équilatéral indirect, il faut passer par l'équateur des triangles plats. Mais cette représentation de l'espace de triangles est beaucoup plus profonde qu'un simple moyen mnémotechnique. Avant tout, elle rend parfaitement compte de l'analyse procustéenne sur les triangles, comme indiqué par le théorème suivant.

Théorème 5.1 (Sphère de Kendall et moindres carrés)**.** *Soit ABC et DEF deux triangles du plan aux sommets étiquetés, identifiés à des triplets d'affixes complexes $z^\star = (z_1^\star, z_2^\star, z_3^\star)$ et $w^\star = (w_1^\star, w_2^\star, w_3^\star)$.*

Suivant la discussion des pages précédentes, on associe à $z^\star$ (respectivement $w^\star$) un couple de complexes "barycentriques" $z = (z_1, z_2)$ (resp. $w = (w_1, w_2)$), puis un ratio projectif complexe $\xi = z_2/z_1$ (resp. $\xi' = w_2/w_1$), un couple d'angles $(\theta, \varphi) \in [0, \pi] \times [0, 2\pi[$ (resp. (θ', φ')) et un point de la sphère centrée en 0 de rayon $1/2$ dans $\mathbb{R}^3$, $f(\xi)$ (resp. $f(\xi')$).

Alors en notant $W = \sqrt{|w_1|^2 + |w_2|^2}$ l'échelle de DEF, on a :

$$\mathrm{d}_{Procuste}(ABC \to DEF) \;=\; \min_{\tau,\upsilon} \|S_{\tau,\upsilon}(z^\star) - w^\star\|_2 \;=\; W \cdot \|f(\xi) - f(\xi')\|_{\mathbb{R}^3}\,. \tag{5.20}$$

Autrement dit : la distance procustéenne entre triangles est directement proportionnelle à la distance euclidienne entre points de la sphère. L'ensemble des triangles du plan peut donc être vu comme le produit de deux facteurs : le groupe des similitudes "$S_{\tau,\upsilon}$" (de dimension réelle 4), et la sphère de Kendall *de dimension 2, qui caractérise les résiduels après recalage rigide.*

La preuve repose sur un simple calcul de minimum par annulation de la dérivée, avec utilisation de moult identités trigonométriques. Pour permettre au plus grand nombre d'accéder à ce résultat, je vais maintenant détailler tous les points de la preuve, sans omettre une ligne de calcul.

Démonstration. Avant tout, reprenons les membres de notre équation. À gauche et au centre, un réel défini de manière implicite :

$$\mathrm{d}_{\mathrm{Procuste}}(ABC \to DEF) \;=\; \min_{\tau,\upsilon\in\mathbb{C}} \sqrt{\sum_{i=1}^{3} |\upsilon \cdot z_i^\star + \tau - w_i^\star|^2}. \tag{5.21}$$

À droite, une distance :

$$W \cdot \|f(\xi) - f(\xi')\|_{\mathbb{R}^3} \tag{5.22}$$

$$= \; W \cdot \left\| \frac{1}{2}(\cos\theta, \sin\theta\cos\varphi, \sin\theta\sin\varphi) - \frac{1}{2}(\cos\theta', \sin\theta'\cos\varphi', \sin\theta'\sin\varphi') \right\|_{\mathbb{R}^3} \tag{5.23}$$

$$= \; W \cdot \frac{1}{2}\sqrt{(\cos\theta - \cos\theta')^2 + (\sin\theta\cos\varphi - \sin\theta'\cos\varphi')^2 + (\sin\theta\sin\varphi - \sin\theta'\sin\varphi')^2} \tag{5.24}$$

où W, θ, θ', φ et φ' sont définis à partir de $z^\star$ et $w^\star$.

Pour identifier ces deux quantités, on procède en deux temps : d'abord, on explicite le terme de gauche pour aboutir à une fonction de z et w ; ensuite, on développe cette expression dans les coordonnées sphériques (θ, φ) jusqu'à voir apparaître le terme de droite.

Passage aux distances quadratiques Commençons par rappeler que $x \mapsto \sqrt{x}$ est une fonction positive et strictement croissante de x. Il suffit donc de démontrer notre égalité sur les carrés des distances mises en jeu, i.e. montrer que :

$$\min_{\tau,\upsilon\in\mathbb{C}} \left(\sum_{i=1}^{3} |\upsilon \cdot z_i^\star + \tau - w_i^\star|^2 \right) \tag{5.25}$$

$$= W^2 \cdot \frac{1}{4}\big((\cos\theta - \cos\theta')^2 + (\sin\theta\cos\varphi - \sin\theta'\cos\varphi')^2 + (\sin\theta\sin\varphi - \sin\theta'\sin\varphi')^2\big) \tag{5.26}$$

Calcul du minimum procustéen Explicité à gauche de l'équation précédente, $\|S_{\tau,\upsilon}(z^\star) - w^\star\|_2^2$ est une fonction positive et continue des variables complexes τ et υ, à $z^\star$ et $w^\star$ fixés. Or on voit qu'elle diverge vers $+\infty$ lorsque $|\tau|$ ou $|\upsilon|$ s'éloigne à l'infini. Par un raisonnement de compacité déjà utilisé équation (2.18) (preuve directe du théorème fondamental de l'algèbre), on sait donc qu'il existe un couple de paramètres optimaux (τ_0, υ_0) qui réalise le minimum.

La similitude optimale recale les barycentres Reste à le caractériser. Pour cela, remarquons que pour tout accroissement $\mathrm{d}\tau$, on a :

$$\|S_{\tau_0+\mathrm{d}\tau,\upsilon_0}(z^\star) - w^\star\|_2^2 \;\geqslant\; \|S_{\tau_0,\upsilon_0}(z^\star) - w^\star\|_2^2 . \tag{5.27}$$

Or on peut calculer que :

$$\|S_{\tau_0+\mathrm{d}\tau,\upsilon_0}(z^\star) - w^\star\|_2^2 \;=\; \sum_{i=1}^{3} |\upsilon_0 \cdot z_i^\star + \tau_0 + \mathrm{d}\tau - w_i^\star|^2 \tag{5.28}$$

$$= \sum_{i=1}^{3} (\upsilon_0 \cdot z_i^\star + \tau_0 + \mathrm{d}\tau - w_i^\star)\,\overline{(\upsilon_0 \cdot z_i^\star + \tau_0 + \mathrm{d}\tau - w_i^\star)} \tag{5.29}$$

$$= \|S_{\tau_0,\upsilon_0}(z^\star) - w^\star\|_2^2 \;+\; 3\,|\mathrm{d}\tau|^2 \tag{5.30}$$

$$+ \sum_{i=1}^{3} \mathrm{d}\tau \cdot \overline{(\upsilon_0 \cdot z_i^\star + \tau_0 - w_i^\star)} \;+\; \overline{\mathrm{d}\tau} \cdot (\upsilon_0 \cdot z_i^\star + \tau_0 - w_i^\star) \tag{5.31}$$

$$= \|S_{\tau_0,\upsilon_0}(z^\star) - w^\star\|_2^2 \;+\; 3\,|\mathrm{d}\tau|^2 \tag{5.32}$$

$$+ \;2\,\mathrm{Re}\left(\mathrm{d}\tau \cdot \sum_{i=1}^{3} \overline{(\upsilon_0 \cdot z_i^\star + \tau_0 - w_i^\star)} \right), \tag{5.33}$$

qui est le développement en $\mathrm{d}\tau$ du coût minimisé au voisinage de l'optimum (τ_0, υ_0). Par suite de l'inégalité (5.27), le terme linéarisé d'ordre 1 doit être positif pour toute valeur assez faible de $\mathrm{d}\tau$: ceci n'est possible que si on a *annulation de la dérivée*,

$$\sum_{i=1}^{3} (\upsilon_0 \cdot z_i^\star + \tau_0 - w_i^\star) \;=\; 0. \tag{5.34}$$

Autrement dit, la condition de minimisation en τ_0 s'écrit à l'ordre 1 :

$$\upsilon \cdot \frac{1}{3}\sum_{i=1}^{3} z_i^\star + \tau_0 = \frac{1}{3}\sum_{i=1}^{3} w_i^\star, \tag{5.35}$$

i.e. « S_{τ_0,υ_0} recale les deux barycentres ». C'est un résultat que nous avions utilisé de manière intuitive dans la Figure 5.3b.

Passage aux coordonnées barycentriques On a vu aux équations (5.12-5.13) que l'on pouvait écrire la dissimilarité procustéenne en coordonnées barycentriques :

$$\|S_{\tau,\upsilon}(z^\star) - w^\star\|_2^2 \;=\; \|S_{\tau,\upsilon}(z) - w\|_2^2 \tag{5.36}$$

$$\;=\; |\upsilon \cdot z_0 + \sqrt{3}\,\tau - w_0|^2 \;+\; |\upsilon \cdot z_1 - w_1|^2 \;+\; |\upsilon \cdot z_2 - w_2|^2. \tag{5.37}$$

Or nous venons de montrer que la similitude optimale recale les barycentres :

$$\upsilon_0 \cdot z_0 + \sqrt{3}\,\tau_0 = w_0. \tag{5.38}$$

Pour trouver le coût procustéen $\mathrm{d}^2_{\mathrm{Procuste}}(ABC \to DEF)$, il suffit donc de minimiser en υ l'expression :

$$\|S_{\tau_0,\upsilon}(z^\star) - w^\star\|_2^2 \;=\; |\upsilon \cdot z_1 - w_1|^2 \;+\; |\upsilon \cdot z_2 - w_2|^2. \tag{5.39}$$

Un coût explicite Comme au paragraphe précédent, on écrit :

$$\|S_{\tau_0,\upsilon_0+\mathrm{d}\upsilon}(z^\star) - w^\star\|_2^2 \;=\; \sum_{i=1}^{2}((\upsilon + \mathrm{d}\upsilon) \cdot z_i - w_i) \cdot \overline{((\upsilon + \mathrm{d}\upsilon) \cdot z_i - w_i)} \tag{5.40}$$

$$\;=\; \|S_{\tau_0,\upsilon_0}(z^\star) - w^\star\|_2^2 \;+\; |\mathrm{d}\upsilon|^2 \cdot (|z_1|^2 + |z_2|^2) \tag{5.41}$$

$$+\; \sum_{i=1}^{2} \mathrm{d}\upsilon \cdot z_i \cdot \overline{(\upsilon \cdot z_i - w_i)} + \overline{\mathrm{d}\upsilon \cdot z_i} \cdot (\upsilon \cdot z_i - w_i)\,. \tag{5.42}$$

La condition d'annulation de la dérivée en υ_0 est donc :

$$\sum_{i=1}^{2} \overline{z_i} \cdot (\upsilon_0 \cdot z_i - w_i) \;=\; 0, \qquad \text{i.e.} \qquad \upsilon_0 \cdot \sum_{i=1}^{2} |z_i|^2 \;=\; \sum_{i=1}^{2} \overline{z_i} w_i. \tag{5.43}$$

À l'optimum, on a alors :

$$\sum_{i=1}^{2} \overline{(\upsilon_0 z_i)} \cdot (\upsilon_0 z_i) \;=\; \sum_{i=1}^{2} \overline{(\upsilon_0 z_i)} \cdot w_i, \tag{5.44}$$

ce qui nous permet de conclure :

$$\mathrm{d}^2_{\mathrm{Procuste}}(ABC \to DEF) = \sum_{i=1}^{2} \overline{(\upsilon_0 z_i - w_i)} \cdot (\upsilon_0 z_i - w_i) \tag{5.45}$$

$$= \sum_{i=1}^{2} \overline{(\upsilon_0 z_i)} \cdot (\upsilon_0 z_i) - \overline{(\upsilon_0 z_i)} \cdot w_i - \overline{w_i} \cdot (\upsilon_0 z_i) + \overline{w_i} \cdot w_i \tag{5.46}$$

$$= \sum_{i=1}^{2} \overline{w_i} \cdot (w_i - \upsilon_0 z_i) \tag{5.47}$$

$$= \sum_{i=1}^{2} \overline{w_i} w_i - \upsilon_0 \sum_{i=1}^{2} \overline{w_i} z_i \tag{5.48}$$

$$= \sum_{i=1}^{2} \overline{w_i} w_i - \frac{|\sum_{i=1}^{2} \overline{w_i} z_i|^2}{\sum_{i=1}^{2} |z_i|^2} \tag{5.49}$$

$$= \left(\sum_{i=1}^{2} |w_i|^2\right) \cdot \left(1 - \frac{|\sum_{i=1}^{2} \overline{w_i} z_i|^2}{\left(\sum_{i=1}^{2} |z_i|^2\right) \cdot \left(\sum_{i=1}^{2} |w_i|^2\right)}\right). \tag{5.50}$$

Un coût fonction des ratios complexes ξ et ξ' À l'aide de deux conditions d'annulation de la dérivée, nous avons pu résoudre notre problème de minimisation : trouver la similitude optimale pour recaler ABC sur DEF n'était finalement pas si difficile.

Maintenant, nous allons chercher à exprimer le coût irréductible du résiduel donné équation (5.50) en fonction des coordonnées sphériques (θ, φ) et (θ', φ'). On a convenu de se doter de ratios ξ et ξ' tels que :

$$z_2 = \xi\, z_1 \qquad \text{et} \qquad w_2 = \xi'\, w_1. \tag{5.51}$$

On peut donc écrire :

$$\mathrm{d}^2_{\text{Procuste}}(ABC \to DEF) = \left(\sum_{i=1}^{2} |w_i|^2\right) \cdot \left(1 - \frac{|\sum_{i=1}^{2} \overline{w_i} z_i|^2}{\left(\sum_{i=1}^{2} |z_i|^2\right) \cdot \left(\sum_{i=1}^{2} |w_i|^2\right)}\right) \tag{5.52}$$

$$= \big(|w_1|^2 + |w_2|^2\big) \cdot \left(1 - \frac{|\overline{w_1} z_1 + \overline{w_2} z_2|^2}{(|z_1|^2 + |z_2|^2) \cdot (|w_1|^2 + |w_2|^2)}\right) \tag{5.53}$$

$$= \big(|w_1|^2 + |w_2|^2\big) \cdot \left(1 - \frac{|\overline{w_1} z_1 + \overline{w_1} z_1 \overline{\xi'} \xi|^2}{|z_1|^2(1 + |\xi|^2) \cdot |w_1|^2(1 + |\xi'|^2)}\right) \tag{5.54}$$

$$= \big(|w_1|^2 + |w_2|^2\big) \cdot \left(1 - \frac{|1 + \overline{\xi'} \xi|^2}{(1 + |\xi|^2) \cdot (1 + |\xi'|^2)}\right). \tag{5.55}$$

En notant $W^2 = |w_1|^2 + |w_2|^2$ la variance du triangle DEF, facteur d'échelle, on trouve :

$$\frac{1}{W^2}\, \mathrm{d}^2_{\text{Procuste}}(ABC \to DEF) \;=\; 1 - \frac{|1 + \overline{\xi'} \xi|^2}{(1 + |\xi|^2) \cdot (1 + |\xi'|^2)}. \tag{5.56}$$

En coordonnées sphériques Pour conclure, il suffit alors de s'armer de patience, d'une fiche de formules trigonométriques, et de se souvenir que

$$\xi \;=\; \tan(\theta/2)\, e^{i\varphi}, \qquad\qquad \xi' \;=\; \tan(\theta'/2)\, e^{i\varphi'}. \tag{5.57}$$

On trouve :

$$\frac{1}{W^2}\, \mathrm{d}^2_{\text{Procuste}}(ABC \to DEF) \;=\; 1 - \frac{|1 + \tan(\theta/2) \tan(\theta'/2) e^{i(\varphi - \varphi')}|^2}{\big(1 + \tan^2(\theta/2)\big) \cdot \big(1 + \tan^2(\theta'/2)\big)} \tag{5.58}$$

$$= \frac{\tan^2(\theta/2) + \tan^2(\theta'/2) - 2 \tan(\theta/2) \tan(\theta'/2) \cos(\varphi - \varphi')}{\big(1 + \tan^2(\theta/2)\big) \cdot \big(1 + \tan^2(\theta'/2)\big)} \tag{5.59}$$

$$= \frac{1}{2} \cdot (1 - \cos\theta \cos\theta' - \sin\theta \sin\theta' \cos(\varphi - \varphi')) \tag{5.60}$$

$$= \frac{1}{4} \cdot \big((\cos\theta - \cos\theta')^2 + (\sin\theta \cos\varphi - \sin\theta' \cos\varphi')^2 + (\sin\theta \sin\varphi - \sin\theta' \sin\varphi')^2\big) \tag{5.61}$$

$$= \|f(\xi) - f(\xi')\|^2_{\mathbb{R}^3}. \tag{5.62}$$

En passant aux racines carrées, on a bien retrouvé l'énoncé du théorème :

$$\mathrm{d}_{\text{Procuste}}(ABC \to DEF) \;=\; W\, \|f(\xi) - f(\xi')\|_{\mathbb{R}^3}. \tag{5.63}$$

À la variance du triangle d'arrivée près, la sphère de rayon 1/2 est bien le bon espace pour calculer les distances procustéennes entre triangles. □

Statistiques sur la sphère

Le Théorème 5.1 confère à la sphère de Kendall un statut *géométrique* privilégié, en montrant qu'elle capture parfaitement la notion de distance (procustéenne) entre triangles. Cette "canonicité" rejaillit sur la structure intrinsèque de la sphère (distance géodésique, surface, symétries) qui se retrouve tout à coup légitimée : elle n'est pas le fruit du choix arbitraire des formules de l'équation (5.17), mais bien du passage au quotient par les similitude de la distance "des moindres carrées", canonique et structurée sur l'espace des coordonnées de points.

Toutes ces propriétés restent néanmoins cantonnées aux seules questions de distances, de géométrie pure. Alors, quelle surprise de découvrir le théorème *probabiliste* suivant :

Théorème 5.2 (Sphère de Kendall et lois normales)**.** *Supposons que les sommets $z_1^\star$, $z_2^\star$ et $z_3^\star$ soient tirés de manière indépendante selon une même loi* normale *(ou* gaussienne*)* $\mathcal{N}((\mu_x, \mu_y); \Sigma)$ *centrée autour d'un point (μ_x, μ_y) avec une matrice de covariance Σ. On va considérer les triangles $z^\star$ à translation, rotation et homothétie près : quitte à faire un changement de repère rigide, on peut donc supposer que $(\mu_x, \mu_y) = (0, 0)$ et que la covariance de la loi s'écrit*

$$\Sigma = \begin{pmatrix} s^2 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \text{avec } s \geqslant 1, \tag{5.64}$$

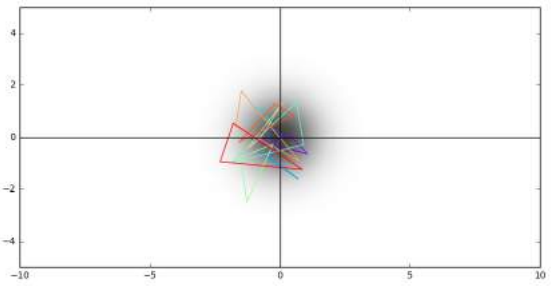
ce qui revient à dire que $x_i^\star = Re(z_i^\star)$ et $y_i^\star = Im(z_i^\star)$ suivent des lois gaussiennes centrées indépendantes d'écarts-types respectifs s et 1.

Alors la variable aléatoire $f(z^\star)$ suit une loi sur la sphère qui est fonction ***de la seule altitude Z*** *et du paramètre d'anisotropie s.*
Si $s = 1$ (loi gaussienne isotrope), on retrouvera exactement la ***loi uniforme*** *sur la sphère.*
Sinon, on observera une concentration autour de l'équateur de triangles plats qui est d'autant plus importante que s est grand.

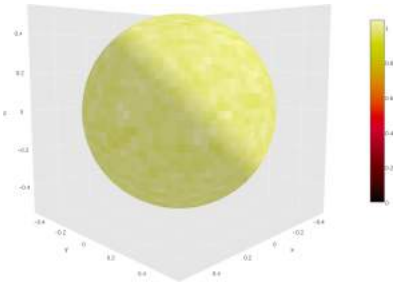
Le résultat ci-dessus est remarquable : il fait le lien entre la loi normale isotrope dans le plan (qui est en un sens la loi de probabilité la plus simple, la moins structurée dans $\mathbb{R}^2$) et la loi uniforme sur la sphère. Autrement dit, voir les triangles sur la sphère est aussi naturel que de tirer des sommets indépendamment selon une loi normale isotrope. Sur la sphère, chaque élément de surface est d'une importance proportionnée à sa représentativité, ce qui est un progrès flagrant par rapport au plan des ξ présenté Figure 5.9.

Intérêt pratique du résultat Si nous tirons les points du plan non plus selon une loi normale, mais de manière uniforme dans un domaine fixé – par exemple, la péninsule de Land's End dans les Cornouailles –, la répartition des formes de triangles sur la sphère s'en trouve sensiblement modifiée. À la dernière ligne de la Figure 5.11, on représente les densités empiriques sur la sphère pour des points tirés uniformément dans le disque unité, puis dans des ellipses de rapports d'anisotropie 2 et 4 pour 1. Ici, point de distribution uniforme sur la sphère : là où la loi normale et sa "bosse" maintenaient à flot les triangles "à deux petits côtés + un grand", ceux-ci sont maintenant clairement désavantagés au profit des triangles à "deux grands côtés + un petit".

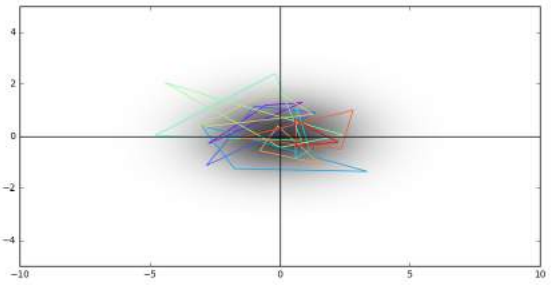
Par contre, le phénomène de tassement sur l'équateur des triangles plats se confirme à mesure que la loi de tirage devient anisotrope. In fine, la comportement précis de la distribution des angles d'un triangle en fonction de la loi de tirage de ses sommets peut être bien comprise, et des tables mises à dispositions des archéologues et des biologistes. Tester une hypothèse d'indépendance dans la génération des sommets par cette statistique est donc devenu une opération de routine.



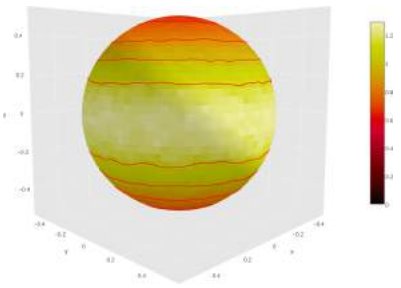
(a) 10 triangles, $s = 1$.



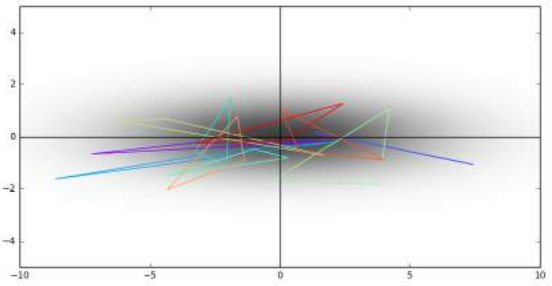
(b) 10^7 triangles, $s = 1$.



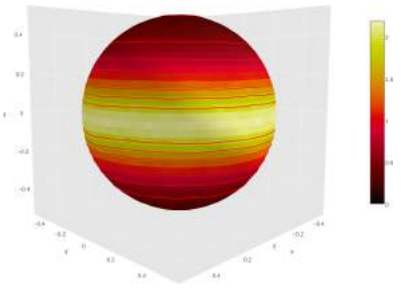
(c) 10 triangles, $s = 2$.



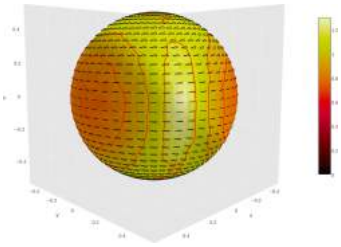
(d) 10^7 triangles, $s = 2$.



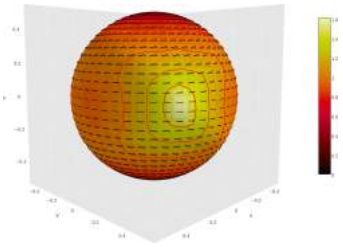
(e) 10 triangles, $s = 4$.



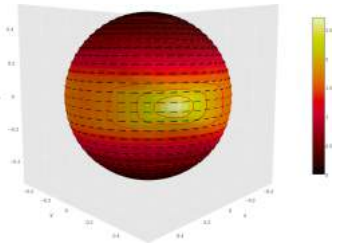
(f) 10^7 triangles, $s = 4$.



(g) 10^7 triangles, $r = 1$.



(h) 10^7 triangles, $r = 2$.



(i) 10^7 triangles, $r = 4$.

FIGURE 5.11 – Histogrammes empiriques sur la sphère de Kendall. Sur les trois premières lignes (a-f), les points sont tirés de manière i.i.d. selon une gaussienne isotrope (a-b) ou anisotrope (c-d) et (e-f). À gauche, on représente la densité de tirage des points et une dizaine de triangles. Comme indiqué par le théorème 5.2, la densité sur la sphère (représentée à droite par rapport à la mesure de surface) ne dépend que de l'altitude Z.
Sur la ligne du bas (g-i), des histogrammes analogues sont tracés pour des points tirés de manière uniforme dans une ellipse de grands axes $(r, 1)$. Les triangles isocèles "deux petits côtés + un grand" perdent en influence au profit des triangles "deux grands côtés + un petit", ce qui conduit à la formation de trois îlots d'attraction espacés régulièrement le long de l'équateur.

Conclusion

Dans ce chapitre, nous avons présenté la théorie des similitudes sous un angle pratique, celui des recalages *rigides*. Nous avons vu comment factoriser la variabilité d'une population d'images $P^1, \ldots, P^N$ en un sous-espace de déformations bien comprises (les *similitudes* $S_{\tau,\upsilon}$), et un espace de *résiduels après recalage* difficiles à interpréter.

Dans le cas des espaces de triangles, le travail de David Kendall nous permet de penser aux résiduels comme aux points d'une sphère : une représentation élégante, pertinente à la fois d'un point de vue *métrique* et *probabiliste*. Il s'agit d'un travail remarquable... Mais difficile à étendre : aucun résultat analogue n'existe pour des espaces de résiduels génériques et le miracle de la sphère des triangles ne semble pas pouvoir être renouvelé.

Alors, comment faire pour aller plus loin, pour analyser de manière *fine* des populations de formes *quelconques* ? La solution, que nous exposerons dans l'ultime chapitre de ce cours, sera d'*élargir* l'espace des déformations admissibles : en enrichissant le groupe des déformations du plan pour dépasser les 4 dimensions des similitudes, on rendra accessibles à l'analyse numérique des variations de formes plus complexes que les seules translations, rotations et changements d'échelle. Tirant parti des idées développées dans tout ce polycopié, il s'agira finalement de faire de la *géométrie riemannienne sur l'espace des déformations fluides du plan.*

Chapitre 6

Un domaine de recherche actuel : l'anatomie computationnelle

Séance 7

Au delà des similitudes : les déformations fluides

Le chapitre précédent nous a permis de nous familiariser avec l'analyse *procustéenne* qui, étant donné deux formes X et Y (disons des nuages de points étiquetés, pour faire simple), permet de trouver une similitude optimale f telle que

$$X \xrightarrow{f} f(X) \leftrightarrows Y, \tag{6.1}$$

$$\text{avec une dissimilarité} \qquad \|f(X) - Y\|_2^2 = \sum_{i=1}^{n} \big(f(X^i) - Y^i\big)^2 \qquad \text{minimale.} \tag{6.2}$$

Par ce biais, l'analyse procustéenne découpe la variabilité $X \to Y$ en un *recalage rigide* $X \to f(X)$ et un *résiduel* $f(X) \to Y$, séparation facile de réaliser puisque l'espace des similitudes est bien connu, de dimension finie – voir Figure 5.3.

Si le résiduel est généralement d'une grande complexité, tout l'intérêt de la méthode repose dans la *description simple* que l'on peut faire du recalage optimal f : on peut *lire* sur les paramètres de f les différences de position, d'échelle et d'orientation qui existent entre X et Y. Cela peut par exemple être utile si l'on cherche à quantifier la répartition des *corpulences* dans une population de poissons : il suffira de recaler toutes les données entre elles, et de regarder la distribution des *échelles de dilatation*.

Malheureusement, une description aussi simple de la variabilité ne saurait suffire. Quid en effet des variations de pose ? Des différences anatomiques plus fines, des rapports relatifs entres les tailles des organes ? On a vu Figure 5.5c que, pour étudier les différences de formes (volume du ventre...) à similitude près, les biologistes avaient eu recours à une étude "en plaques minces" sur les résiduels $Y - f(X)$: de quoi s'agit-il au juste ?

Un modèle de déformation simpliste – celui des similitudes – ne permet d'extraire qu'une quantité d'information *limitée* – position, échelle, orientation. Pour aller plus loin, transférer plus de renseignements du *résiduel* au *recalage* f, il va nous falloir *assouplir* les conditions qui portent sur ce dernier.

N.B. : Le cœur du chapitre est rédigé de manière extrêmement technique, à la façon d'un support d'exposé entre doctorants. Pour terminer ce polycopié, je tenais à vous montrer ce à quoi peut ressembler un véritable dialogue entre mathématiciens. Il est bien clair qu'en cours, je m'attacherai à vous présenter ce travail d'une manière accessible à tous !

Un point de vue radicalement opposé : le transport optimal

En contrepoint de l'analyse procrustéenne, on va maintenant présenter une autre méthode pour *recaler* une forme sur une autre : celle qui repose sur la théorie du *transport* de Gaspard Monge (1746-1818) et Leonid Kantorovitch (1912-1986).

Rappelons-le : la théorie du recalage rigide développée au chapitre 5 modélisait chaque forme par un nuage de points *étiquetés à l'avance*, et permettait de calculer des *similitudes optimales* pour passer d'une image à l'autre. À l'inverse, la théorie du transport s'intéresse au cas où *aucun étiquetage* a priori n'est disponible : nos images sont des dessins ou des photos "brutes", ni plus ni moins.

Des images aux mesures Soit X une image, que l'on modélisera par une courbe polygonale donnée d'une collection de segments du plan $([a^i, b^i])_{1 \leqslant i \leqslant I}$. Pour conserver des calculs raisonnables, on va procéder à une simplification brutale : l'oubli de la structure *topologique* de la courbe, pour se concentrer sur la seule *masse* du dessin, en remplaçant cette collection de segments par une *mesure*

$$\mu = \text{Mesure}(X) = \sum_{i=1}^{I} \mu_i \delta_{x^i}, \tag{6.3}$$

somme de petits diracs (masses ponctuelles) de masses μ_i localisés aux points x^i, où

$$x^i = \frac{a^i + b^i}{2} \qquad \text{et} \qquad \mu_i = \left\| b^i - a^i \right\| \tag{6.4}$$

sont respectivement les centres et les longueurs des segments de X.

Transport Suivant les idées de Monge, recaler une image X sur une image Y, c'est *transporter* la "masse d'encre"

$$\mu = \text{Mesure}(X) = \sum_{i=1}^{I} \mu_i \delta_{x^i} \qquad \text{sur la mesure} \qquad \nu = \text{Mesure}(Y) = \sum_{j=1}^{J} \nu_j \delta_{y^j}. \tag{6.5}$$

Mais comment s'y prendre, au juste ? En 1781, date à laquelle il publie son *Mémoire sur la théorie des déblais et des remblais*, Monge a en tête des problèmes de défense nationale, de construction de places fortes. L'unité de masse est pour lui la motte de terre ou le sac de sable, et il s'agit de transporter une masse totale M tirée de fossés localisés en les points x^i sur les fondations des remblais en y^j. D'un côté, I sources fournissant chacune μ_i mottes de terre aux points x^i ; de l'autre, J points à renforcer de ν_j mottes aux points y^j. On a bien sûr :

$$\sum_{i=1}^{I} \mu_i \;=\; M \;=\; \sum_{j=1}^{J} \nu_j \qquad \text{avec des masses que l'on supposera entières, pour simplifier.} \tag{6.6}$$

Formulation du problème Pour Monge, il s'agit de travailler vite et bien ; d'économiser autant que faire se peut la sueur des ouvriers, tout en achevant le travail dans la journée. Si on convient que les instants $t = 0$ et $t = 1$ correspondent respectivement au début et à la fin de la journée de travail, il s'agit pour lui d'*optimiser* les trajets des M mottes de terre au cours des travaux. Pour tout indice m dans $[\![1, M]\!]$, on notera

$$\gamma^m : \; t \in [0,1] \;\mapsto\; \gamma_t^m \in \mathbb{R}^2 \tag{6.7}$$

la trajectoire de la m^{e} brouette, et on conviendra d'un coût de déplacement *quadratique* : entre t et $t + \mathrm{d}t$, transporter la brouette d'une position γ_t^m à $\gamma_{t+\mathrm{d}t}^m = \gamma_t^m + \dot\gamma_t^m \,\mathrm{d}t$ occasionnera un effort "$\|\dot\gamma_t^m\|^2 \,\mathrm{d}t$" d'autant plus important que le déplacement est rapide. (En son temps, Monge utilisait un simple coût linéaire "$\|\dot\gamma_t^m\| \,\mathrm{d}t$" qui se révéla être plus difficile à analyser.)

Le problème de transport *continu* est donc le suivant :

$$\text{Trouver les trajectoire } \gamma_m \text{ qui minimisent} \qquad \sum_{m=1}^{M} \int_{t=0}^{1} \|\dot\gamma_t^m\|^2 \,\mathrm{d}t \tag{6.8}$$

sous la contrainte que, pour tous indices $i \in [\![1, I]\!]$ et $j \in [\![1, J]\!]$,

$$\#\big\{m \in [\![1, M]\!] \ , \ \gamma_0^m = x^i\big\} \ = \ \mu_i, \tag{6.9}$$

$$\#\big\{m \in [\![1, M]\!] \ , \ \gamma_1^m = y^j\big\} \ = \ \nu_j. \tag{6.10}$$

Des brouettes au plan de transport A priori, le problème ci-dessus est extrêmement difficile à résoudre puisqu'il porte sur des vecteurs de dimension infinie, les chemins γ^m. Mais heureusement, le coût (6.8) est si simple que l'on peut pré-optimiser chaque chemin indépendamment des autres ; le Théorème 4.2 sur les géodésiques du plan euclidien (les lignes droites) permet en effet d'affirmer qu'à extrémités γ_0^m et γ_1^m fixées, γ_t^m est entièrement déterminé :

$$\gamma_t^m : \ t \in [0,1] \ \mapsto \ (1-t)\cdot\gamma_0^m + t\cdot\gamma_1^m, \tag{6.11}$$

avec un coût

$$\int_{t=0}^{1} \|\dot\gamma_t^m\|^2 \,\mathrm{d}t \ = \ \|\gamma_1^m - \gamma_0^m\|^2 . \tag{6.12}$$

On peut alors remarquer que le coût total n'est plus fonction que d'un grand *plan* global $\Gamma = (\gamma_{i,j})_{(i,j)\in[\![1,I]\!]\times[\![1,J]\!]}$ déterminant combien de mottes de terre doivent être envoyées de x^i à y^j en ligne droite, pour un coût $\gamma_{i,j} \left\|x^i - y^j\right\|^2$. En notant $c_{i,j} = \left\|x^i - y^j\right\|^2$ le coût du déplacement d'un sac de x^i vers y^j, on trouve la *formulation statique du problème de transport* :

$$\text{Trouver le plan } \Gamma \text{ qui minimise} \qquad \sum_{i,j} \gamma_{i,j}\, c_{i,j}, \tag{6.13}$$

sous la contrainte que :

$$\forall\ i,j,\ \ \gamma_{i,j} \geqslant 0, \qquad \forall\ i,\ \ \sum_j \gamma_{i,j} = \mu_i, \qquad \forall\ j,\ \ \sum_i \gamma_{i,j} = \nu_j. \tag{6.14}$$

Le transport optimal est un problème d'étiquetage Finalement, le transport optimal n'est plus qu'un problème d'*affectation*. Si on a numéroté les mottes de terre au départ de 1 à M et qu'on a fait de même pour les emplacements d'arrivée, on peut encoder le plan de transport Γ par une permutation (i.e. une *bijection*) $\sigma : [\![1, M]\!] \to [\![1, M]\!]$ et on trouve un *coût de transport* :

$$C^{X,Y}(\sigma) \ = \ \sum_{m=1}^{M} \left\|x^m - y^{\sigma(m)}\right\|^2 . \tag{6.15}$$

Ici, les masses μ_i et ν_j sont rendues par le fait que plusieurs mottes de terres peuvent se trouver au même endroit.

Calcul efficace de plans de transports diffus Calculer ces transports en un temps raisonnable a longtemps été un problème : comme tout problème d'assignement *combinatoire*, il était ardu à résoudre de manière exacte. Heureusement, depuis 2013 et la publication par Marco Cuturi de l'article *Sinkhorn Distances : Lightspeed Computation of Optimal Transport*, on dispose d'un algorithme itératif dit de *Sinkhorn* pour calculer extrêmement rapidement des plans de transports *diffus*, *probabilistes*, qui approximent bien l'étiquetage déterministe optimal.

Procuste ou Monge ? Le *transport optimal* met l'accent sur une notion de déformation *économique* et flexible, là où l'analyse procustéenne priorisait la *rigidité* des recalages obtenus. En pratique, les biologistes, neurologues et médecins ont besoin de recalages intermédiaires : plus souples que les similitudes, mais moins irréguliers que les plans de transport simples. Comment les obtenir ? C'est tout l'objet des pages qui suivent.

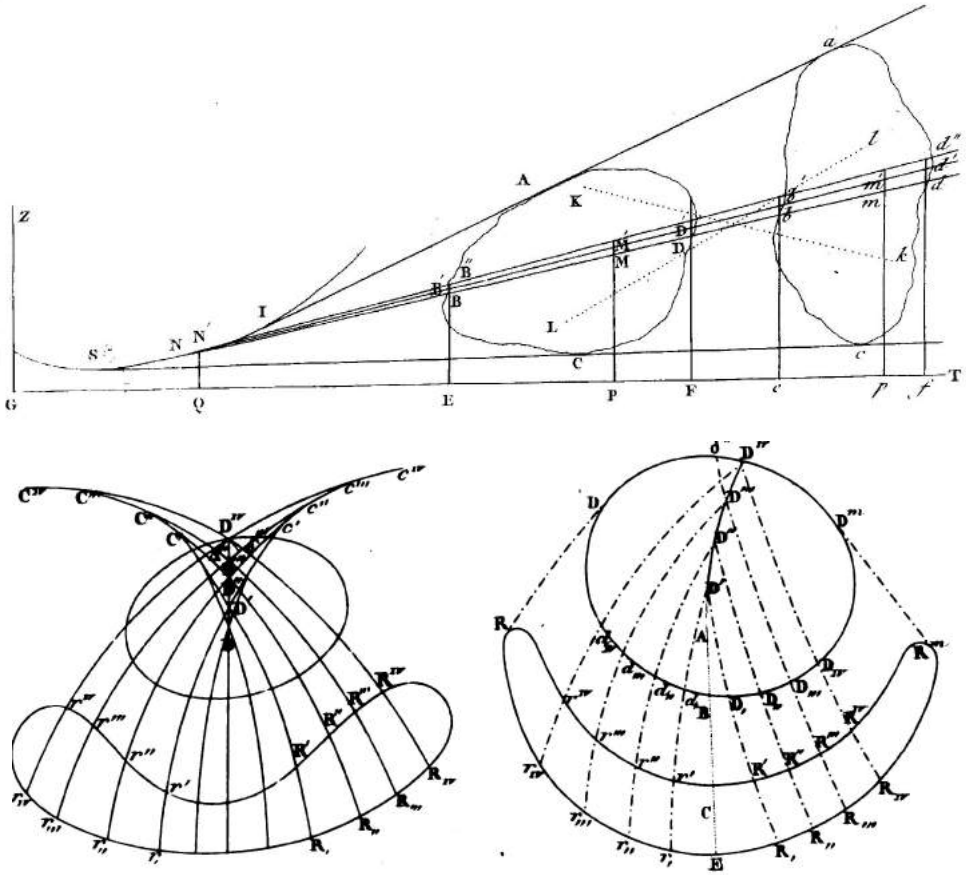
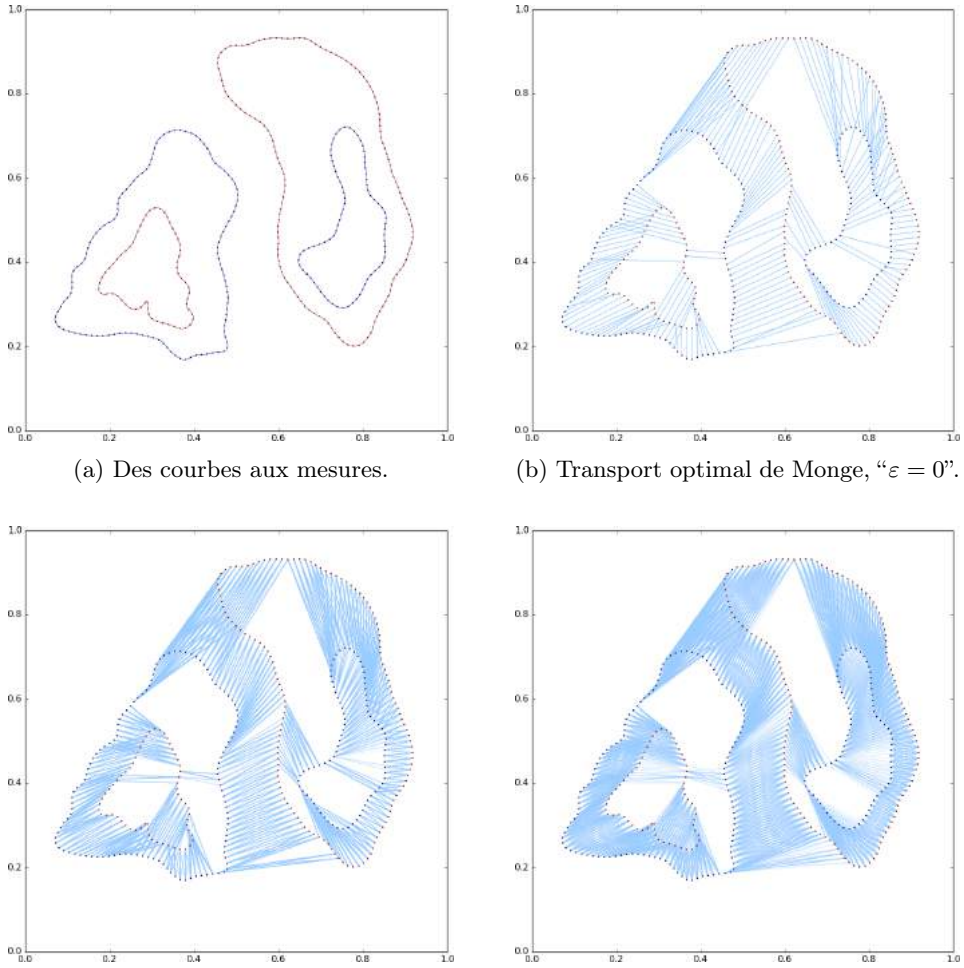


FIGURE 6.1 – Mon petit exposé sur le transport optimal est bourré d'anachronismes... Pour retrouver l'esprit du mémoire de Monge, n'hésitez pas à lire l'article *Le mémoire sur les déblais et les remblais* d'Étienne Ghys, disponible sur Images des Maths : `images.math.cnrs.fr/Gaspard-Monge,1094.html`.



(a) Des courbes aux mesures.

(b) Transport optimal de Monge, "$\varepsilon = 0$".

(c) Transport diffus, $\sqrt{\varepsilon} = .01$.

(d) Transport diffus, $\sqrt{\varepsilon} = .03$.

FIGURE 6.2 – Plans de transport optimaux entre une forme bleue (les déblais) et une courbe rouge (les remblais), vues dans le plan $\mathbb{R}^2$ (une carte d'état major). Première étape de l'analyse : le passage des *courbes* aux *mesures* de masse (a). Pour simplifier le problème, on normalise les deux masses totales pour obtenir le même nombre de "mottes de terre" au départ et à l'arrivée. Le plan de transport optimal (b) peut alors être calculé par un algorithme *combinatoire* ; mieux : l'algorithme de *Sinkorn* permet d'obtenir en un temps réduit des approximations *probabilistes* du transport qui dépendent du seul paramètre de diffusion ε.
Si cette théorie permet d'obtenir rapidement des correspondances point-à-point entre formes, elle n'est pas adaptée aux problèmes d'imagerie *médicale*. Imaginez seulement les conséquences d'un tel *matching* si courbes bleues et courbes rouges représentaient des coupes du cœur (avec deux ventricules...) à deux instants différents du cycle cardiaque !
C'est qu'en oubliant d'entrée de jeu toute information *topologique* sur les voisinages et les continuités de la forme, la modélisation de Monge et Kantorovitch jette aux oubliettes une information *cruciale*.

Une régularisation qui passe par la géométrie Riemannienne

Pour limiter les déchirures, une idée pertinente semble être de *régulariser le transport*. Encourager deux points proches au *départ* à s'envoyer sur des points proches à l'*arrivée*. On peut par exemple chercher un plan de transport $\sigma : [\![1, M]\!] \to [\![1, M]\!]$ qui minimise un coût "régularisé"

$$\mathcal{C}_k^{X,Y}(\sigma) = \underbrace{\sum_m \left\| x^m - y^{\sigma(m)} \right\|^2}_{\text{Coût du transport}} + \underbrace{\sum_{m,m'} k(x^m, x^{m'}) \cdot \left\| y^{\sigma(m)} - y^{\sigma(m')} \right\|^2}_{\text{Coût de régularisation}}, \tag{6.16}$$

où la fonction $k : (x^m, x^{m'}) \mapsto \exp(-\|x^m - x^{m'}\|^2 / 2\, l^2)$ (par exemple) est une fonction de *corrélation* qui donne un poids élevé (ici, jusqu'à 1) aux couples de points de départ $(x^m, x^{m'})$ tels que l'écart $\|x^m - x^{m'}\|$ est inférieur à une échelle caractéristique l, et qui n'accorde aucune importance aux autres. Ainsi, un transport σ optimal au sens de $\mathcal{C}_k^{X,Y}$ serait *régulier* en plus d'être *paresseux* : si deux points sont k-proches au départ (i.e. avec $k(x^m, x^{m'})$ non négligeable), alors ils ne sauraient s'envoyer sur des points éloignés à l'arrivée puisque l'écartement est pénalisé de manière quadratique en $\|y^{\sigma(m)} - y^{\sigma(m')}\|^2$.

Vers une interprétation géométrique du coût de transport Si séduisante que puisse sembler la formule 6.16, elle souffre d'une grave limitation théorique : elle ne conduit pas à une *distance* entre formes. En effet, étant donné un coût de transport $\mathcal{C}_k$, nous aimerions pouvoir définir la distance "transport" entre deux nuages de points X et Y comme la plus petite valeur possible du coût $\mathcal{C}_k(\sigma)$:

$$\mathrm{d}_k(X, Y) = \min_\sigma \ \mathcal{C}_k^{X,Y}(\sigma). \tag{6.17}$$

Une telle distance pourrait nous permettre de plaquer toutes nos intuitions géométriques sur les espaces de formes mis en jeu, à condition de vérifier trois axiomes fondamentaux :

Séparation : $\mathrm{d}_k(X, Y) = 0 \iff X = Y$ à la numérotation près.
Symétrie : $\mathrm{d}_k(X, Y) = \mathrm{d}_k(Y, X)$
Inégalité triangulaire : $\mathrm{d}_k(X, Z) \leqslant \mathrm{d}_k(X, Y) + \mathrm{d}_k(Y, Z)$

On peut vérifier que la distance engendrée par le transport "classique" vérifie bien ces axiomes : elle est appelée "distance de Wasserstein", et fait toujours l'objet de nombreuses études. Par contre, la distance "naïve" associée à la formule (6.16) ne vérifie ni l'axiome de symétrie, ni l'inégalité triangulaire. Elle est donc à peu près inutilisable en tant qu'outil d'analyse.

Le problème de la formule (6.16) est le rôle privilégié qu'elle accorde aux positions de départ x^m, qui sont seules utilisées comme "indicateurs de voisinages" au travers de la fonction k. Pour rattraper la sauce, on peut symétriser la formule par l'ajout d'un deuxième terme de régularisation,

$$\mathcal{C}_{k,\mathrm{sym}}^{X,Y}(\sigma) = \underbrace{\sum_m \left\| x^m - y^{\sigma(m)} \right\|^2}_{\text{Coût du transport}} + \frac{1}{2} \underbrace{\sum_{m,m'} k(x^m, x^{m'}) \cdot \left\| y^{\sigma(m)} - y^{\sigma(m')} \right\|^2}_{\text{Coût de régularisation } X \to Y} \tag{6.18}$$

$$+ \frac{1}{2} \underbrace{\sum_{m,m'} k(y^m, y^{m'}) \cdot \left\| x^{\sigma^{-1}(m)} - x^{\sigma^{-1}(m')} \right\|^2}_{\text{Coût de régularisation } Y \to X}. \tag{6.19}$$

Malheureusement, la "distance" engendrée par un tel coût ne tient compte que des formes d'arrivée et de départ, sans considération pour les positions des porteurs au cours du trajet.

D'une affectation atomique à un transport continu Pour concevoir une notion de transport régularisé qui reste intuitive, il nous faut revenir au problème de Monge initial, à l'interprétation *cinématique* du transport. Étant donnée une trajectoire $\gamma : t \mapsto \gamma_t = (\gamma_t^1, \dots, \gamma_t^M)$ telle que $\gamma_0^m = x^m$ et $\gamma_1^m = y^{\sigma(m)}$, on avait défini le coût du transport

$$\mathcal{C}(\gamma) = \int_0^1 \sum_m \|\dot\gamma_t^m\|^2 \,\mathrm{d}t, \tag{6.20}$$

qui se trouvait valoir

$$\mathcal{C}^{X,Y}(\sigma) = \sum_m \left\| x^m - y^{\sigma(m)} \right\|^2 \tag{6.21}$$

pour un transport optimal en ligne droite des x^m aux $y^{\sigma(m)}$.

Pour accéder à un transport optimal corrélé *bien fondé*, avec une interprétation cinématique naturelle qui permette à terme de garantir les axiomes de la distance, il conviendra de revenir à une pénalisation des trajets, de modifier la formule (6.20) en

$$\mathcal{C}_k(\gamma) = \int_0^1 \Bigg[\underbrace{\sum_m \|\dot\gamma_t^m\|^2}_{\text{Coût du transport}} + \underbrace{\sum_{m,m'} k(\gamma_t^m, \gamma_t^{m'}) \cdot \left\| \dot\gamma_t^m - \dot\gamma_t^{m'} \right\|^2}_{\text{Coût de régularisation}} \Bigg] \mathrm{d}t. \tag{6.22}$$

Cela revient à remplacer nos transporteurs indépendants – brouettes de terre ou caisses de munitions – par des unités interdépendantes, grégaires, qui marchent à l'unisson lorsqu'elles sont k-proches, répugnant à s'éloigner les unes des autres. Des particules de miel, par exemple.

Le transport corrélé est un problème Riemannien A priori, l'équation (6.22) impose une pénalisation compliquée, à deux termes, sur le transport de M particules corrélées entre elles. **Comment, alors, réussir à trouver des chemins $\mathcal{C}_k$-optimaux ?**

Pour y répondre, un mathématicien commencera par chercher des simplifications conceptuelles. On préférera voir cette équation comme un problème de transport simple sur *l'espace des n-uplets de points* muni d'une métrique Riemannienne arbitraire. Ainsi, considérons l'espace de landmarks

$$\mathcal{L}_M^2 = \left\{ (q^1, \dots, q^M) \in \mathbb{R}^2 \ , \ q^i \neq q^j \right\} \tag{6.23}$$

des M-uplets de points distincts deux à deux dans le plan $\mathbb{R}^2$. On peut le voir comme un ouvert de l'espace $\mathbb{R}^{M\times 2}$, et décrire un petit voisinage de tout nuage $q = (q^1, \dots, q^M)$ comme l'ensemble des nuages $q + v$, où $v = (v^1, \dots, v^M)$ est un M-uplet de vecteurs suffisamment petits.

De la même manière que le disque de Poincaré était muni d'un champ de températures qui dilatait les distances, on peut munir l'espace des landmarks d'une métrique Riemannienne $q \mapsto g_q$ donnée par

$$\frac{\left(\mathrm{d}_g(q \to q + v \cdot \mathrm{d}t)\right)^2}{\mathrm{d}t} = \sum_m \|v^m\|^2 + \sum_{m,m'} k(q^m, q^{m'}) \cdot \left\| v^m - v^{m'} \right\|^2 \tag{6.24}$$

$$= \quad v^\mathsf{T} g_q v \quad = \quad \|v\|_{g_q}^2 \tag{6.25}$$

pour tout nuage de points $q = (q^1, \dots, q^M) \in \mathcal{L}_M^2$ et tout champ de vitesses $v = (v^1, \dots, v^M) \in \mathbb{R}^{M\times 2}$. Ici, g_q est une matrice $(M \times 2)$-par-$(M \times 2)$ symétrique définie positive, une *métrique* qui pénalise les déviations de points de q proches au sens de la fonction k.

Retour sur le coût de transport On peut dire qu'on a muni l'espace des nuages de point d'une métrique *qui pénalise les déchirures*, les séparations de points proches au sens d'une fonction noyau k. Surtout, on peut maintenant réécrire le coût

$$\mathcal{C}_k(\gamma) = \int_0^1 \|\dot{\gamma}_t\|^2_{\gamma_t} \, \mathrm{d}t \tag{6.26}$$

et l'on s'est ramené à chercher, dans l'espace Riemannien $(\mathcal{L}^2_M, g_q)$, le(s) chemin(s) optimal entre deux formes X et Y, i.e. un chemin γ dans l'espace des landmarks qui réalise le minimum du coût $\mathcal{C}_k(\gamma)$ sous les conditions

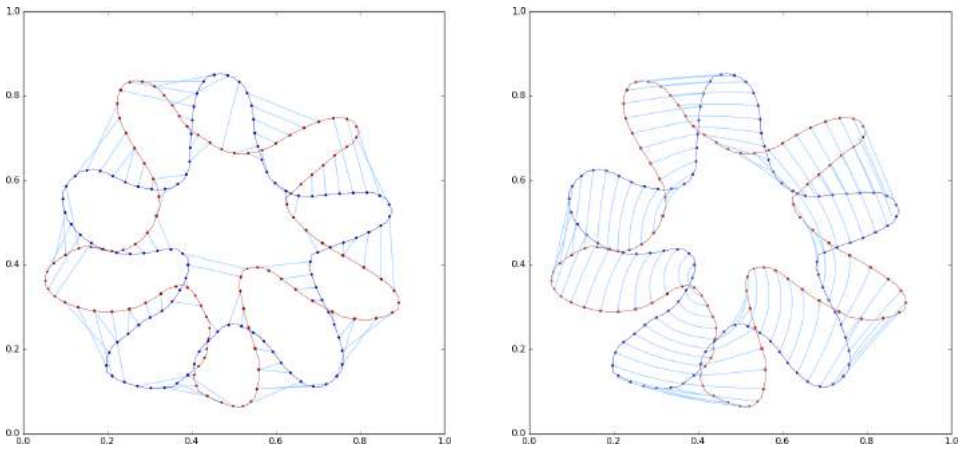
$$\gamma_0 = X, \qquad \gamma_1 = Y. \tag{6.27}$$

Formulé ainsi, notre problème est analogue à celui du pilote d'avion devant aller de Paris à Moscou en une heure, à moindre frais : il s'agit de joindre un point à un autre sur une variété courbe, en minimisant la consommation de carburant. Avec un coût quadratique en la norme de la vitesse, le trajet optimal sera celui qui parcourt la distance de X à Y sur un trajet de longueur minimale, *à vitesse constante* : plutôt que d'être lièvre, de se reposer 30mn pour courir ensuite deux fois plus vite sur les 30 minutes restantes, on aura toujours intérêt à se faire tortue, à avancer d'un pas égal.

Finalement, le problème du transport régularisé peut être reformulé comme suit :

« Quel est le chemin géodésique minimisant la longueur entre X et Y,
dans l'espace des nuages de points munis de la métrique anti-déchirures g_q ? » (6.28)

À une question compliquée posée dans un espace simple, on a substitué une question simple dans un espace compliqué. C'est une idée fructueuse, car qui dit question simple dit méthode, et qui dit méthode dit résolution !



(a) Par le transport optimal. (b) Par une méthode riemannienne.

FIGURE 6.3 – Un avant-goût des bénéfices apportés par une méthode riemannienne de *régularisation* du transport. Il s'agit ici de trouver une trajectoire *optimale* d'un nuage de points bleu X vers un nuage rouge Y. À gauche (a), la trajectoire de type *transport* (pourtant calculée avec une possibilité d'effacement des *outliers* trop importants) n'est pas satisfaisante : elle découpe les bras de l'étoile de mer sans considérations pour sa topologie.
Ce problème sera résolu par l'utilisation de métriques régularisantes sur l'espace des nuages de points (b), au prix d'un surcoût algorithmique conséquent (perte de convexité de la fonctionnelle minimisée, etc.).

Tir géodésique sur une variété Riemannienne

Au chapitre 4, nous avons pu décrire explicitement les géodésiques de trois espaces homogènes : le plan euclidien, la sphère et le plan hyperbolique (ou disque de Poincaré). Grâce aux nombreuses isométries de ces espaces, nous avions en effet réussi à nous ramener à des cas simples, accessibles au calcul et au raisonnement direct.

En toute généralité, ce n'est plus possible : la Figure 6.4 montre bien que sur une surface quelconque, on ne peut espérer de description exhaustive simple des courbes géodésiques. D'ailleurs, ce n'est pas parcequ'une courbe est *géodésique*, ou "droite" (ce qui est une notion locale : entre deux points proches, elle minimise la distance) qu'elle est nécessairement un plus court chemin entre ses deux extrémités : on peut souvent trouver d'autres géodésiques plus astucieuses, qui coupent à travers les "vallées" !

Équations géodésiques Heureusement, il est tout de même possible d'obtenir quelques résultats au sujet de ces "lignes droites". Si on considère une courbe $\gamma : t \mapsto \gamma_t$ à valeurs dans une variété Riemannienne $(\mathcal{M}, g)$ de dimension D (disons, une surface plongée dans l'espace ambiant, ou nos espaces de nuages de points), le fait d'être "droit" est assez restrictif et impose une forme de conservation de la "direction".

Quitte à reparamétrer γ (ce qui ne change pas le trajet effectué), on peut en fait montrer que γ est une géodésique si et seulement si elle vérifie "l'équation des géodésiques avec symboles de

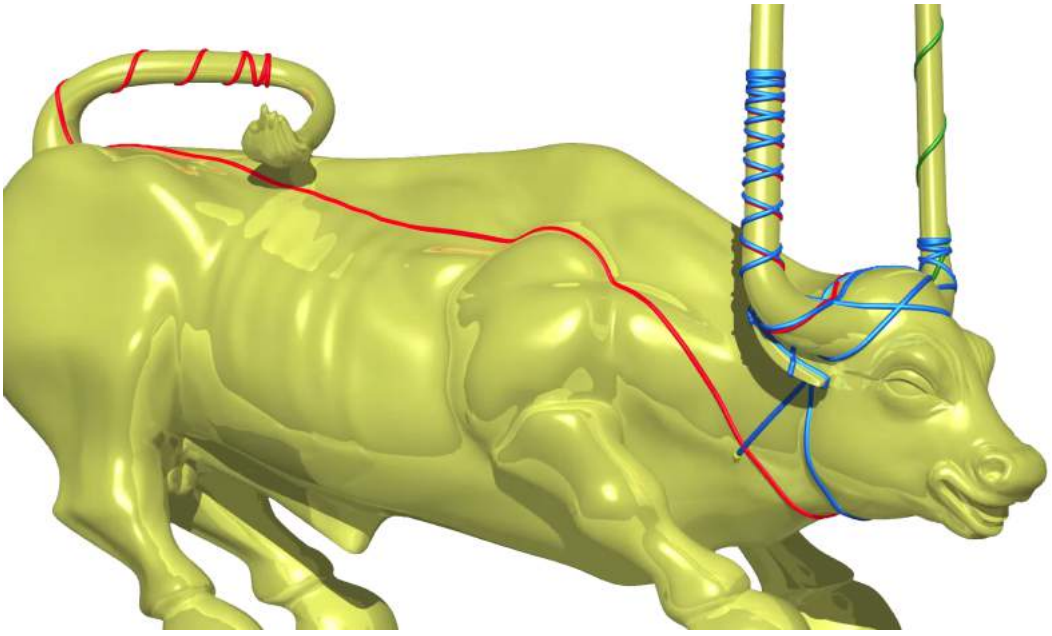


FIGURE 6.4 – Trois exemples de trajectoires géodésiques – i.e. qui minimisent localement la longueur – sur une statue de taureau. On l'aura compris, l'ensemble des "lignes droites" est ici bien plus riche que sur la sphère homogène !
Image tirée du très bon film *Chaos* de Jos Leys, Étienne Ghys et Aurélien Alvarez, accessible librement à l'adresse suivante : `www.chaos-math.org`.

Christoffel" : dans un système de coordonnées locales $\gamma_t = (\gamma_t^1, \dots, \gamma_t^D)$,

$$\forall\ d \in [\![1, D]\!]\,,\quad \ddot{\gamma}_t^d = -\sum_{i,j=1}^{D} \Gamma_{ij}^d(\gamma_t) \cdot \dot{\gamma}_t^i\, \dot{\gamma}_t^j, \tag{6.29}$$

où le symbole de Christoffel $\Gamma_{ij}^d(q)$ est donné par :

$$\Gamma_{ij}^d(q) = \frac{1}{2}\sum_{l=1}^{D} g^{dl}(q) \cdot (\partial_i g_{jl}(q) + \partial_j g_{il}(q) - \partial_l g_{ij}(q))\,, \tag{6.30}$$

avec g_{ij} les coefficients de la métrique, et g^{dl} ceux de son inverse, la cométrique. L'accélération d'une particule géodésique γ_t est donc fonction lisse de sa position et de sa vitesse, avec un contrôle quadratique en cette dernière.

Équations Hamiltoniennes du mouvement La description ci-dessus permet d'assurer que pour tout couple position-vitesse (q, v), ***il existe une unique trajectoire géodésique* sur la variété passant en $t = 0$ par le point q avec la vitesse v.** C'est un bon début, conséquence du Théorème B.4.2 (Cauchy-Lipschitz), mais il semble difficile d'aller plus loin : l'équation 6.29 manque de structure, et est trop générique pour pouvoir être vraiment travaillée au corps.

Heureusement, un changement de variables astucieux (des vitesses v vers les moments p) va nous tirer d'affaire et nous permettre de continuer notre étude. C'est l'objet du théorème suivant :

Théorème 6.1 (Caractérisation Hamiltonienne des géodésiques d'une variété Riemannienne)**.** *Soit $\gamma : t \mapsto \gamma_t$ une courbe lisse à valeurs dans une variété Riemannienne $(\mathcal{M}, g)$, que l'on peut supposer de vitesse de norme constante quitte à la reparamétrer. Notons*

$$q_t = \gamma_t, \qquad\qquad p_t = g_{q_t}(\dot{q}_t) \tag{6.31}$$

qui sont respectivement la position *et le* moment *associés à la courbe au temps t. Le moment p_t n'est rien d'autre que la vitesse $v_t = \dot{\gamma}_t$ vue au travers d'un changement de variable adéquat donné par la co-métrique :*

$$K_{q_t} = (g_{q_t})^{-1}. \tag{6.32}$$

À un point (q, p) de l'espace des phases (le fibré cotangent), on associe l'énergie cinétique correspondante via le Hamiltonien :

$$H(q, p) = \frac{1}{2} p^\mathsf{T} K_q p, \tag{6.33}$$

de sorte qu'à chaque instant, on ait :

$$\frac{1}{2}\|v_t\|_{g_{\gamma_t}}^2 = \frac{1}{2} v_t^\mathsf{T} g_{q_t} v_t = \frac{1}{2} p_t^\mathsf{T} K_{q_t} p_t = H(q_t, p_t). \tag{6.34}$$

On a alors équivalence entre les deux propositions suivantes :

1. *γ est une courbe* géodésique*, qui minimise (localement) la distance.*
2. *La trajectoire relevée (q_t, p_t) dans l'espace des phases suit le flot hamiltonien correspondant au gradient symplectique ("tourné de 90°") de H :*

$$(\text{Ham}) : \left\{ \begin{array}{rcccl} \dot{q}_t & = & +\frac{\partial H}{\partial p}(q_t, p_t) & = & +K_{q_t} p_t \\ \dot{p}_t & = & -\frac{\partial H}{\partial q}(q_t, p_t) & = & -\partial_q(p_t, K_q p_t)(q_t) \end{array} \right. . \tag{6.35}$$

Liens avec la mécanique Le résultat précédent est d'une importance historique *capitale*. Découvert en 1833 par Hamilton, il a révolutionné la physique pour devenir au tournant du XX$^{\text{e}}$ la base des trois grandes théories de la mécanique : quantique, classique, relativiste.

Et pour cause. Dans la section B.4.2, nous avons montré comment formaliser les équations de la mécanique Newtonienne portant sur une particule de masse m en une équation différentielle ordinaire d'ordre 1 sur le couple position-vitesse $(q(t), v(t))$:

$$\begin{pmatrix} \dot{q}(t) \\ \dot{v}(t) \end{pmatrix} = \begin{pmatrix} v(t) \\ \frac{1}{m} F(q(t), v(t)) \end{pmatrix}, \tag{6.36}$$

$$\text{i.e.} \qquad \ddot{q}(t) = \frac{1}{m} F(q(t), v(t)) \tag{6.37}$$

avec $F(q(t), v(t))$ la somme des forces s'exerçant sur la particule, fonction uniquement de sa position (interaction gravitationnelle, ressort...) et de sa vitesse (force de Lorentz en électromagnétisme). Il s'agit du parfait analogue de l'équation des géodésiques (6.29).

Dans l'espace des phases position-vitesse (q, v) – aussi appelé *fibré tangent* –, obéir aux lois de Newton revient à suivre un flot lisse,

$$X(q, v) = \begin{pmatrix} v \\ \frac{1}{m} F(q, v), \end{pmatrix}. \tag{6.38}$$

En conséquence, le théorème de Cauchy-Lipschitz peut garantir le *déterminisme* des lois de Newton : étant donné un couple position-vitesse (q, v) en $t = 0$, il existe une unique trajectoire physique (q_t, v_t), définie pour tout instant réel t et obéissant aux lois de la mécanique. La connaissance exacte des positions et vitesses initiales à un instant donné permet donc de décrire entièrement le passé, le présent et le futur d'un système mécanique.

L'espace des phases position-moment Le résultat d'Hamilton permet d'aller *beaucoup* plus loin dans l'analyse grâce à une idée géniale : l'abandon de la vitesse v au profit du moment $p = (g_q)^{-1} v$ – les mécaniciens utilisent une autre définition plus physique, mais équivalente.

Il ne s'agit a priori que d'un changement de variables linéaires, d'un changement de coordonnées locales dans l'espace des vitesses. On a du mal à en saisir l'intérêt a priori : pourquoi délaisser la vitesse v au profit d'une grandeur ad hoc ? C'est que de manière remarquable, les équations du mouvement/des géodésiques (c'est la même chose) prennent une forme toute particulière, symétrisée dans le nouvel espace des phases positions-moments (q, p) : en dénotant $X(q, p)$ le flot associé à l'équation (6.35), on a

$$X(q, p) = \begin{pmatrix} +\frac{\partial H}{\partial p}(q, p) \\ -\frac{\partial H}{\partial q}(q, p) \end{pmatrix} = \text{“R}_{-90^\circ}\text{”}\big(\nabla H(q, p)\big). \tag{6.39}$$

Petit exercice : le montrer dans le cas d'une particule d'altitude z, de masse m soumise à un champ gravitationnel d'intensité g. On prendra $q = z$, $v = \dot{q}$, $p = mv$ et

$$H(q, p) = \text{“}E_{\text{mec}}\text{”}(q, p) = \text{“}E_{\text{cin}}\text{”}(q, p) + \text{“}E_{\text{pp}}\text{”}(q, p) = \frac{1}{2}\frac{p^2}{m} + mgq. \tag{6.40}$$

Être une trajectoire géodésique/physique dans l'espace des positions, c'est donc suivre dans l'espaces des phases un simple flot stationnaire, donné par le gradient “tourné à 90°” de l'énergie mécanique (en termes techniques : le gradient symplectique du Hamiltonien).

Contrairement au flot “générique” des équations de Christoffel/Newton, le flot hamiltonien de l'équation (6.39) est parfaitement symétrique en q et p, naturel si on admet le principe d'évolution à énergie mécanique constante. **Au sens de l'évolution géodésique/physique du système, la variable naturelle n'est donc pas la vitesse v, mais bien le moment p.**

Métriques à noyaux, premières intuitions

Rappelons notre motivation initiale : trouver des géodésiques sur l'espace des nuages de points muni de la métrique de transport corrélé "naïve" donnée équation (6.25). Il y a quelques pages encore, cette pénalisation nous semblait être la régularisation la plus naturelle du transport optimal classique, avec une métrique g_q pénalisant les vitesses v de manière extrêmement simple.

La vision Hamiltonienne de l'équation des géodésiques nous force à réviser notre jugement. Dans une optique de recherche de trajectoires optimales, **une métrique n'est pas "naturelle" si elle agit de manière simple sur les *vitesses* : ce qui importe, c'est la manière dont elle pénalise les *moments*.**

Contraintes pratiques sur la cométrique Dans les applications, on l'a vu, il sera primordial de pouvoir calculer des géodésiques de manière efficace. Partant d'une condition initiale (q_0, p_0) (ou (q_0, v_0), de manière équivalente), il s'agira de pouvoir intégrer, *tirer* l'unique géodésique solution de l'équation différentielle (6.35) prenant ces valeurs en $t = 0$. En pratique, on pourra tout simplement le faire via un schéma d'Euler d'ordre 1, en discrétisant l'intervalle de temps en une succession finie d'instants. De $[0, 1]$, on passe à $\{0, 0.1, 0.2, \dots, 1\}$, et on utilise un schéma itératif

$$\text{(Ham. discret)} : \left\{ \begin{array}{rcl} q_{t+0.1} & = & q_t + 0.1 \cdot K_{q_t} p_t \\ p_{t+0.1} & = & p_t - 0.1 \cdot \partial_q (p_t, K_q p_t)(q_t) \end{array} \right. . \tag{6.41}$$

Toute la complexité algorithmique de cette routine de *tir géodésique* se trouve concentrée dans le calcul de la matrice K_{q_t} et du gradient $\partial_q (p_t, K_q p_t)(q_t)$. Pour obtenir des algorithmes *utilisables*, il faut donc nous restreindre à des cométriques K_q **faciles à calculer**.

À une forme $q = (q^1, \dots, q^M) \in \mathcal{L}^2_M$, il s'agit désormais d'associer une grosse matrice symétrique K_q de taille $(M \times 2)$-par-$(M \times 2)$, qui pénalise les moments $p = (p^1, \dots, p^M)$ de manière pertinente. **On en est venu à chercher une application de cométrique $\mathbf{q \mapsto K_q}$ qui soit :**

— Facile à calculer.

— Telle que la métrique $(K_q)^{-1}$ fasse sens, et ressemble à la métrique de transport régularisé.

Métriques à noyaux En matière de simplicité algorithmique, difficile de faire plus efficace que la *cométrique de noyau k*, définie par blocs via :

$$K_q = \left(\begin{array}{c|c|c|c} k(q^1, q^1) I_2 & k(q^1, q^2) I_2 & \cdots & k(q^1, q^M) I_2 \\ \hline k(q^2, q^1) I_2 & k(q^2, q^2) I_2 & \cdots & k(q^2, q^M) I_2 \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline k(q^M, q^1) I_2 & k(q^M, q^2) I_2 & \cdots & k(q^M, q^M) I_2 \end{array} \right), \tag{6.42}$$

avec k une fonction de *noyau* raisonnable, $k(x, y) = 1/(1 + \|x - y\|^2 / l^2)$ par exemple. Il s'agit du produit de Kronecker entre la matrice M-par-M de noyau réduite

$$k_q = \left(\begin{array}{cccc} k(q^1, q^1) & k(q^1, q^2) & \cdots & k(q^1, q^M) \\ k(q^2, q^1) & k(q^2, q^2) & \cdots & k(q^2, q^M) \\ \vdots & \vdots & \ddots & \vdots \\ k(q^M, q^1) & k(q^M, q^2) & \cdots & k(q^M, q^M) \end{array} \right) \tag{6.43}$$

et un bloc identité I_2 de taille 2-par-2. Étant donné un moment $p = (p^1, \dots, p^M)$ et en écrivant $p^i = (p^i_1, p^i_2)$, utiliser la cométrique ci-dessus revient à donner le hamiltonien par la formule

$$H(q, p) \;=\; \frac{1}{2} p^\mathsf{T} K_q p \;=\; \frac{1}{2} \sum_{l=1}^{2} \sum_{i,j=1}^{M} sk(q^i, q^j) \cdot p^i_l \, p^j_l . \tag{6.44}$$

Calcul de la métrique associée sur un exemple jouet Les cométriques à noyaux sont extrêmement simples à calculer : il suffit d'une application de la fonction noyau pour chaque couple (q^i, q^j). La matrice de noyau réduite pourra alors être vue comme une matrice de corrélation entre les points du nuage, qui associe un poids fort aux paires de points proches – le théorème de Mercer, bien connu des spécialistes de l'apprentissage, permet d'ailleurs de préciser cette intuition : k_q est la matrice de Gram d'un plongement non-linéaire des données q^i dans un espace arbitraire.

Pour des applications réelles en imagerie médicale, avec des nuages de plus de 10 000 points en dimension 3, c'est un plus indéniable. Reste à voir si la métrique associée fait sens ! Avant d'attaquer les vrais résultats (qui sont surprenants !), prenons le temps de détailler les calculs dans un cas simple, avec $k(x, y) = \exp(-\|x - y\|^2 / 2\,l^2)$ le noyau gaussien d'échelle l.

On considère un état $q \in \mathcal{L}_6^2$, donné par six points $(q^1, q^2, q^3, q^4, q^5, q^6)$ du plan sur lesquels on fait les hypothèses simplificatrices suivantes :

— Les points se répartissent en trois groupes de taille variable : (q^1), (q^2, q^3, q^4), et (q^5, q^6), qui sont éloignés les uns des autres à une distance très grande devant l'échelle l du noyau.

— q^5 et q^6 sont à une distance d donnée l'un de l'autre.

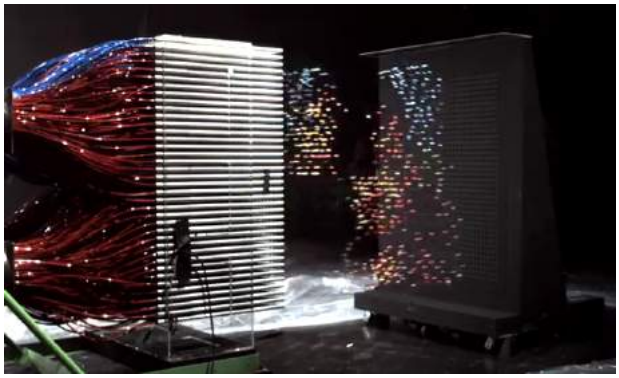
— Le groupe (q^2, q^3, q^4) est un triangle équilatéral de côté d.

On peut voir Figure 6.5 une telle situation, qui modélise de manière simple la présence d'amas de masses variées dans l'image.

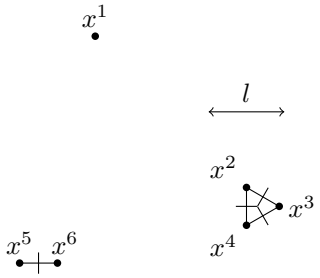
Sous ces hypothèses, on peut écrire très simplement la matrice de noyau réduite :

$$k_q = \begin{pmatrix} 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 1 & a & a & \cdot & \cdot \\ \cdot & a & 1 & a & \cdot & \cdot \\ \cdot & a & a & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & a \\ \cdot & \cdot & \cdot & \cdot & a & 1 \end{pmatrix}, \tag{6.45}$$

où $a = \exp(-d^2/2\,l^2) \in [0, 1]$, et où l'on a remplacé les termes négligeables par des points.



(a) Vidéo *Mythbusters Demo GPU versus CPU*, de la chaîne YouTube Nvidia : allez la regarder !



(b) Nuage jouet de six points.

FIGURE 6.5 – (a) Motivation pratique derrière le choix des métriques à noyaux : la mise à disposition sur le marché de *cartes graphiques*, puces informatiques totalement parallèles qui permettent en un temps court de remplir un tableau de formules simples. (b) Exemple jouet d'un nuage de six points en dimension 2. Les segments marqués sont tous de longueur d, vue au travers du noyau k par le réel $a = \exp(-d^2/2\,l^2) \in [0, 1]$.

À la limite, la matrice de noyau réduite k_q puis son homologue vectoriel K_q sont des matrices diagonales par blocs, remplies de blocs élémentaires

$$B_n(a) = (1-a)\cdot I_n + a\cdot(1)(1)^\mathsf{T} = \begin{pmatrix} 1 & & & & \\ & 1 & & (a) & \\ & & \ddots & & \\ & (a) & & 1 & \\ & & & & 1 \end{pmatrix}. \tag{6.46}$$

Pour trouver la métrique, inverser k_q, la clé est donc de savoir inverser les blocs $B_n(a)$ associés aux amas de n points.

Lemme 6.1 (Pertinence des métriques à noyaux, version discrète)**.** *On note $e = (1)/\|(1)\|_2$ le vecteur unitaire constant de taille n, rempli de $1/\sqrt{n}$. L'inverse de $B_n(a)$ est alors donné par :*

$$(B_n(a))^{-1} = \frac{1}{1+(n-1)\,a} ee^\mathsf{T} + \frac{1}{1-a}(I_n - ee^\mathsf{T}). \tag{6.47}$$

Autrement dit, pour tout vecteur $v = (v^1, \ldots, v^n)$ que l'on décompose en

$$\begin{aligned} v &= e\,(e^\mathsf{T}v) + (v - e\,(e^\mathsf{T}v)) && (6.48)\\ &= v_{moy} + v_{var}\ , && (6.49) \end{aligned}$$

une partie "moyenne" constante et une partie de somme nulle, la variance. On a

$$v^\mathsf{T}(B_n(a))^{-1}v = \frac{1}{1+(n-1)\,a}\,\|v_{moy}\|_2^2 + \frac{1}{1-a}\,\|v_{var}\|_2^2 \tag{6.50}$$

Démonstration. Il suffit d'écrire la décomposition spectrale de $B_n(a)$, i.e. trouver les axes de l'ellipsoïde associé :

$$B_n(a) = (1+(n-1)\,a)\,ee^\mathsf{T} + (1-a)\,(I_n - ee^\mathsf{T}). \tag{6.51}$$

$B_n(a)$ possède donc une valeur propre $1+(n-1)\,a$ selon la direction e, et agit comme $(1-a)$ fois l'identité sur l'orthogonal. Pour trouver l'inverse, il suffit alors d'inverser les valeurs propres – qui correspondent ici aux valeurs singulières, il n'y a vraiment aucun piège. □

Interprétation Rappelons que $a = \exp(-(d/l)^2/2)$, où d est le diamètre de l'amas et l l'échelle du noyau utilisé par la cométrique. Il vaut donc 1 si $d << l$, et décroît jusqu'à 0 lorsque $d >> l$. L'équation (6.50) est extrêmement précieuse, car elle contient en germe tout la dynamique associée à la cométrique K_q.

D'abord, elle met en évidence un fait rassurant : le rôle particulier joué par les champs de vitesses constants, "colinéaires". Le coût associé à un champ de vitesses sur l'amas est donc la somme d'un terme de *translation*, proportionnel à $\|v_{\text{col}}\|_2^2$, et d'un terme de *régularisation* pénalisant la non-uniformité en $\|v_{\text{ncol}}\|_2^2$.

Lorsque le diamètre d de l'amas est bien supérieur à l'échelle du noyau, a est petit devant 1. On a alors

$$\frac{1}{1+(n-1)\,a} \simeq 1 \simeq \frac{1}{1-a}, \tag{6.52}$$

l'équilibre entre les deux pénalisations. Le coût $v^\mathsf{T}(B_n(a))^{-1}v$ est simplement égal au coût Wasserstein $v^\mathsf{T}v = \|v\|_2^2$ du transport décorrélé. On dira que les particules n'*interagissent* pas ensemble.

À l'inverse, si le diamètre d de l'amas est petit devant l, si le noyau voit les points de l'amas comme quasiment confondus, on aura $a \simeq 1^-$ et par suite

$$v^\mathsf{T}(B_n(a))^{-1}v \;\simeq\; \frac{1}{n}\,\|v_{\text{moy}}\|_2^2 \;+\; \infty\,\|v_{\text{var}}\|_2^2\,. \tag{6.53}$$

Lorsque les points sont l-proches les uns des autres, qu'ils interagissent entre eux au sens de k, on a donc combinaison de deux effets : la sur-pénalisation des non-uniformités, des déchirures, avec le poids quasi-infini devant v_{var} ; la mutualisation des coûts de translation, avec une atténuation en $1/n$ du coût quadratique sur v_{moy}. **Tout se passe donc comme si notre amas de n particules se réduisait à un seul atome, très difficile à éclater mais aussi facile à transporter qu'une particule seule.**

Retour sur la forme globale, combinaison de plusieurs amas Si l'on revient au nuage de la Figure 6.5, on peut maintenant exprimer simplement la métrique g_q associée par k_q aux déformations infinitésimales de sa géométrie :

$$g_q = (k_q)^{-1} = \left(\begin{array}{c|ccc|cc} 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \hline \cdot & & & & \cdot & \cdot \\ \cdot & & B_3(a)^{-1} & & \cdot & \cdot \\ \cdot & & & & \cdot & \cdot \\ \hline \cdot & \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & \multicolumn{2}{c}{B_2(a)^{-1}} \end{array}\right) \tag{6.54}$$

(on se dispense ici d'écrire le produit de Kronecker avec I_d, qui impose simplement de sommer les coûts sur les dimensions). Les trois amas sont donc complètement indépendants, ce qui n'est pas une surprise puisqu'ils sont décorrélés au sens de k.

Interprétation Étant donné un champ de vitesses $v = (v^1, v^2, v^3, v^4, v^5, v^6)$, comment se trouve-t-il pénalisé par g_q ? Les comportements limites se retrouvent aussi facilement que pour un amas simple. Aussi, lorsque $d >> l$ et donc $a \simeq 0$, on a :

$$v^\mathsf{T} g_q v \;\simeq\; \|v^1\|_2^2 + \|v^2\|_2^2 + \|v^3\|_2^2 + \|v^4\|_2^2 + \|v^5\|_2^2 + \|v^6\|_2^2. \tag{6.55}$$

Par contre, si $d << l$, alors un champ de coût fini s'écrit nécessairement :

$$(v^1, v^2, v^3, v^4, v^5, v^6) = (w^1, w^2, w^2, w^2, w^3, w^3), \tag{6.56}$$

avec w^i la "vitesse de groupe" de l'amas i, et :

$$\begin{aligned} v^\mathsf{T} g_q v \;&=\; \|v^1\|_2^2 + \frac{1}{3}\big(\|v^2\|_2^2 + \|v^3\|_2^2 + \|v^4\|_2^2\big) + \frac{1}{2}\big(\|v^5\|_2^2 + \|v^6\|_2^2\big) && (6.57)\\ &=\; \|w^1\|_2^2 + \|w^2\|_2^2 + \|w^3\|_2^2. && (6.58) \end{aligned}$$

Cahin-caha, on peut donc se forger une certaine intuition des trajectoires géodésiques "typiques", qui tiennent groupés les k-amas.

Des cométriques à noyaux aux déformations fluides

Les intuitions développées dans les pages précédentes ne concernent a priori que des nuages de points *finis*. Mais de manière particulièrement élégante, celles-ci vont pouvoir être généralisées à la dimension infinie et s'en trouver clarifiées.

Normes RKHS sur les champs de vecteurs du plan Précisons. On se donne pour commencer une fonction de noyau lisse à appliquer aux différences relatives $q^i - q^j$, par exemple $k(x) = \exp(-\|x\|^2 / 2\,\sigma^2)$. Entre autres hypothèses techniques sur le noyau k, liées à la théorie des *espaces à noyaux reproduisant* ("RKHS" en anglais), on supposera que la transformée de Fourier $\widehat{k}$ est réelle, *strictement positive* sur tout le plan fréquentiel. Si $v : \mathbb{R}^2 \to \mathbb{R}^2$ est un champ de vecteurs du plan, on propose de lui assigner une *k-norme* par la formule

$$\|v\|_k^2 \;=\; \int_{\omega\in\mathbb{R}^2} \frac{1}{\widehat{k}(\omega)}\,|\widehat{v}(\omega)|^2\,\mathrm{d}\omega, \tag{6.59}$$

où $\widehat{k}$ et $\widehat{v}$ sont les transformées de Fourier respectives de k et de v dans $\mathbb{R}^2$ – voir le chapitre 3 consacré à cet outil et à ses premières propriétés.

Premiers exemples Si k était très localisé en 0, avec en limite le dirac δ_0 de transformée $\widehat{\delta_0}(\omega) = 1$, on retrouverait la norme L^2 de $\widehat{v}(\omega)$ et, par le théorème 3.1 (Identité de Parseval) et l'équation (3.43), on aurait $\|v\|_k^2 = \|\widehat{v}\|_2^2 = \|v\|_2^2$.

De manière plus pertinente, si k est *lisse*, $\widehat{k}(\omega)$ devient vite très petit à mesure que $|\omega|$ tend vers $+\infty$. La multiplication dans le domaine de Fourier par $1/\widehat{k}(\omega)$ agit donc comme un filtre **passe-haut** d'autant plus puissant que le support fréquentiel de $\widehat{k}$ est localisé.

Intégration d'un champ de vecteurs variable k-lisse In fine, on définit V_k l'ensemble des champs de vecteurs v tels que la k-norme associée $\|v\|_k$ soit finie. Pour peu que k soit suffisamment lisse (ce que l'on suppose désormais), V_k est constitué de champs de vecteurs dont le profil fréquentiel décroît à l'infini plus rapidement que $(\widehat{k}(\omega))^{1/2} \xrightarrow{|\omega|\to\infty} 0$. Aussi, si v_t est un champ de vecteurs variable dépendant du temps tel que

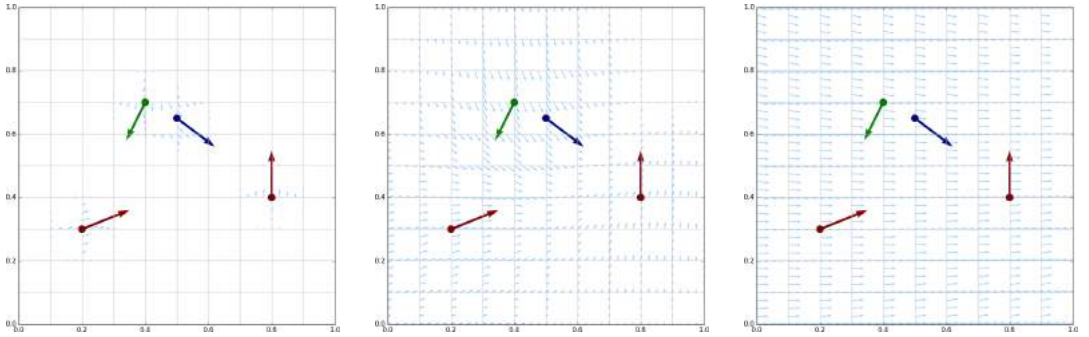
$$\int_0^1 \|v_t\|_k^2\,\mathrm{d}t < +\infty, \tag{6.60}$$

on peut appliquer le Théorème B.4.2 de Cauchy-Lipschitz et s'en servir pour intégrer le *flot de* v_t en une déformation de l'identité : il existe une unique trajectoire φ_t de difféomorphismes du plan telle que :

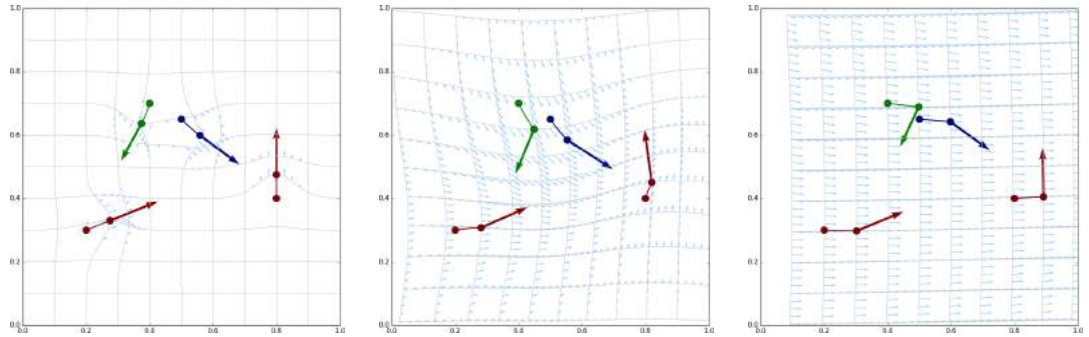
$$\forall\, x \in \mathbb{R}^2,\ \varphi_0(x) = x \qquad \text{et} \quad \forall\, x \in \mathbb{R}^2, \forall\, t \in [0,1],\ \frac{\mathrm{d}}{\mathrm{d}t}\,[\varphi_t(x)] = v_t \circ \varphi_t(x), \tag{6.61}$$

$$\text{i.e.} \qquad \varphi_0 \;=\; \mathrm{Id}_{\mathbb{R}^2} \quad \text{et} \qquad \forall\, t \in [0,1],\ \varphi_t \;=\; \mathrm{Id}_{\mathbb{R}^2} \;+\; \int_{s=0}^{t} v_s \circ \varphi_s\,\mathrm{d}s. \tag{6.62}$$

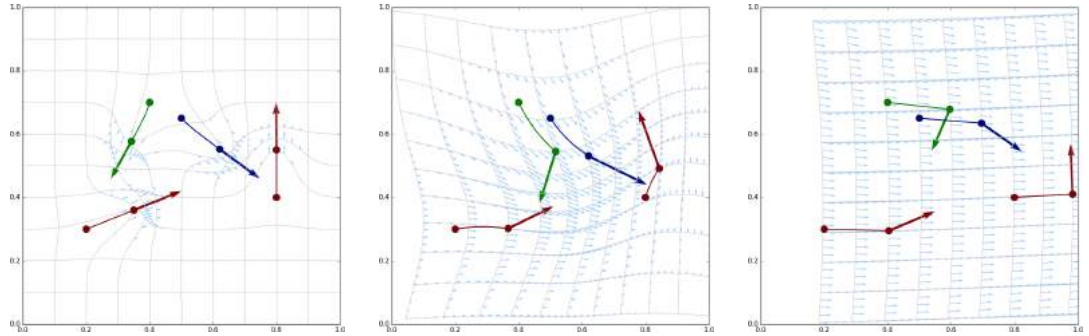
Si φ_0 est identifié à la grille identité, intégrer le flot de v_t entre 0 et T revient à "laisser couler" le plan selon ce courant, et à utiliser la grille déformée comme l'indicateur d'un changement de repère souple φ_T, dont la régularité des v_t permet de garantir le caractère *difféomorphe*.



(a) $\sigma = .25$, $t = 0$. (b) $\sigma = .5$, $t = 0$. (c) $\sigma = 1.$, $t = 0$.



(d) $\sigma = .25$, $t = 1/2$. (e) $\sigma = .5$, $t = 1/2$. (f) $\sigma = 1.$, $t = 1/2$.



(g) $\sigma = .25$, $t = 1$. (h) $\sigma = .5$, $t = 1$. (i) $\sigma = 1.$, $t = 1$.

FIGURE 6.6 – Exemples de trajectoires géodésiques dans l'espace de landmarks $\mathcal{L}_4^2$ des 4-uplets de points, muni de cométriques gaussiennes de variation respectives $\sigma = .25$, $.5$ et 1. Suivant le Théorème 6.2, on peut relever ces trajectoires comme des chemins géodésiques dans les espaces de difféomorphismes du plan G_k associés aux différents rayons.
Ici, les points colorés sont aussi les supports des moments p_t^m, flèches dénotant "l'intention de mouvement" du m^{e} landmark au temps t. On représente les trajectoires difféomorphiques φ_t comme des grilles : valant le quadrillage identité en $t = 0$, elles sont déformées à mesure que les points charrient l'espace aux travers des champs de vitesse $v_t = k \star p_t$, tracés en bleu à petite échelle.
Ces images permettent de comprendre intuitivement l'influence du paramètre d'échelle σ (rayon caractéristique du noyau k) sur les recalages générés par notre théorie : plus σ est grand, plus $\widehat{k}$ tend rapidement vers 0 et plus les difféomorphismes de G_k sont *réguliers*.

Géométrie Riemannienne en dimension infinie Déformer l'espace ambiant par l'écoulement d'un champ de vecteurs variable : une idée qui peut surprendre... Mais qui se place dans la droite ligne du chapitre 4 sur l'étude des chemins dans des variétés riemanniennes.

Pour le comprendre, il suffit de reconnaître (au moins de manière informelle) l'ensemble des difféomorphismes du plan comme une variété de dimension infinie dont le plan tangent est partout identifiable à un espace de champs de vecteurs plus ou moins réguliers – les déformations infinitésimales de la grille.

Dans notre cas, on se restreint à l'ensemble G_k des difféomorphismes obtenus à partir de l'identité par intégration du flot d'un champ de vecteur variable v_t vérifiant la condition d'intégrabilité (6.60) : il s'agit *in fine* d'une variété de dimension infinie modelée sur l'espace V_k dont les points sont des difféomorphismes, et les chemins, des trajectoires $t \mapsto \varphi_t$ dont la dérivée temporelle peut être identifiée à un champ de vecteur k-régulier *variable*.

Géodésiques sur des espaces de difféomorphismes En 1966, Vladimir Arnold, publiait dans les Annales de l'Institut Fourier un article fondateur, *Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfait.* Adoptant un vocabulaire riemannien pour décrire l'ensemble des difféomorphismes de l'espace homotopes à l'identité, il montrait essentiellement le résultat suivant : un flot "Lagrangien" de difféomorphismes $\varphi_t(x)$ obéit à l'équation d'Euler incompressible (i.e. décrit l'écoulement d'un fluide incompressible sans viscosité) si et seulement s'il s'agit d'une géodésique sur l'ensemble des difféomorphismes de l'espace ambiant préservant le volume, muni de la métrique "L^2" – pour nous, informellement, le cas où "$k = \delta_0$".

Cette géométrisation des équations de la mécanique des fluides a stimulé un important travail de recherche en physique, dont un résultat remarquable est le théorème suivant :

Théorème 6.2 (Cométriques à noyaux et déformations de l'espace ambiant, Principe de réduction)**.** *Soit k une fonction de noyau vérifiant toutes les hypothèses de régularité et de positivité de la transformée de Fourier énoncées jusqu'ici.*
Alors toute trajectoire géodésique *q_t dans l'espace des nuages de points $\mathcal{L}^2_M$ muni de la cométrique "K_q" peut se relever en une trajectoire géodésique dans l'espace G_k des k-difféomorphismes du plan muni de la métrique riemannienne $\|\cdot\|^2_k$.*
De manière plus explicite, si (q_t, p_t) est la paramétrisation d'une trajectoire géodésique dans l'espace des phases, alors l'évolution de q_t peut être vue comme le transport du nuage de points q_0 sous l'action d'un continuum de difféomorphismes

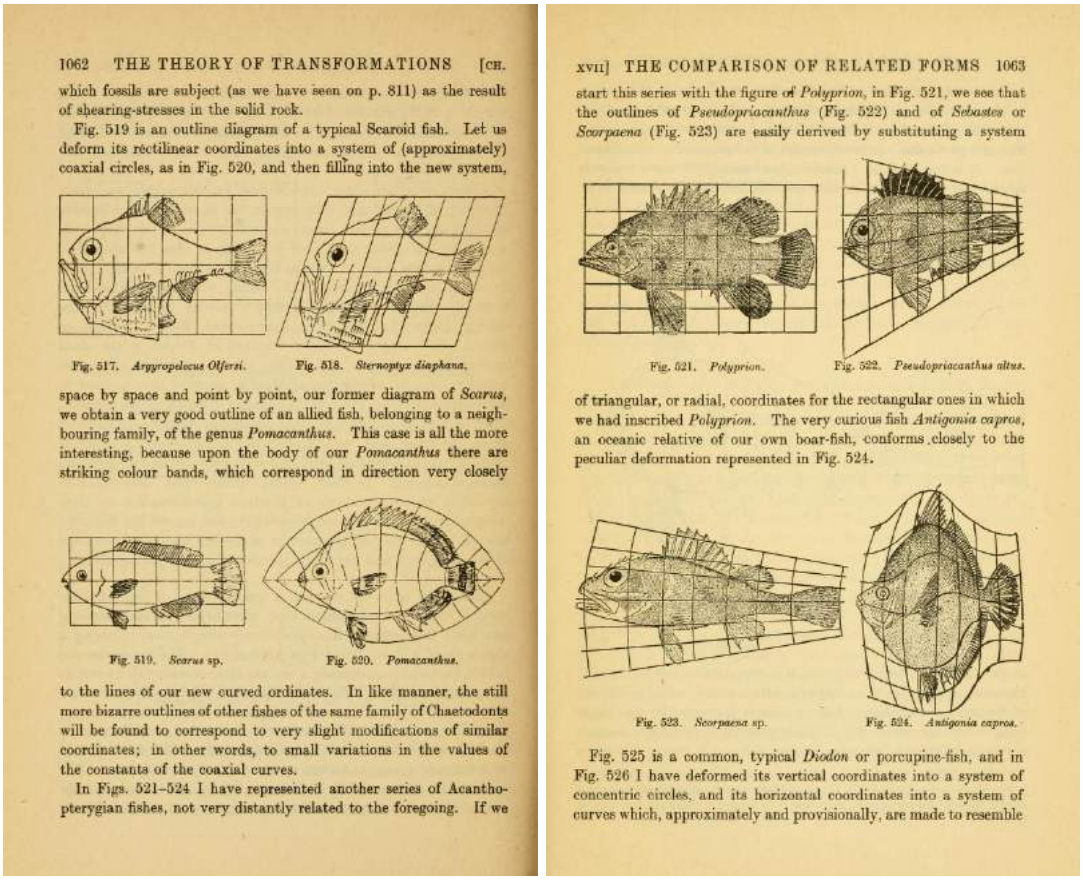
$$\forall\ m \in [\![1, M]\!], \forall\ t, \quad q_t^m \ = \ \varphi_t(q^m), \tag{6.63}$$

où φ_t résulte de l'intégration à partir de l'identité du flot G_k-géodésique

$$v_t(\,\cdot\,) \ = \ \sum_{m=1}^{M} k(\,\cdot\,, q_t^m)\, p_t^m \ = \ k \star p_t, \tag{6.64}$$

en considérant que le "moment d'intention" p_t est une mesure vectorielle portée par les sommets q_t du nuage de points transformé.

Démonstration. La petite analyse matricielle sur les nuages de six points devrait déjà vous avoir convaincu de la pertinence de ce résultat. Une preuve formelle est sans doute superflue dans cet ouvrage de vulgarisation : je me contenterai donc d'esquisser "avec les mains" une preuve élémentaire, par projection orthogonale du champ v_t à chaque instant t sur l'espace des déplacements infinitésimaux "utiles", qui déplacent la forme. Une preuve véritablement moderne, plus facile à généraliser, reposera plutôt sur le *Principe du Maximum de Pontryagin*, résultat fondamental de la théorie du *contrôle optimal*.



1062 THE THEORY OF TRANSFORMATIONS [CH.

which fossils are subject (as we have seen on p. 811) as the result of shearing-stresses in the solid rock.

Fig. 519 is an outline diagram of a typical Scaroid fish. Let us deform its rectilinear coordinates into a system of (approximately) coaxial circles, as in Fig. 520, and then filling into the new system,

Fig. 517. *Argyropelecus Olfersi.* Fig. 518. *Sternoptyx diaphana.*

space by space and point by point, our former diagram of *Scarus*, we obtain a very good outline of an allied fish, belonging to a neighbouring family, of the genus *Pomacanthus*. This case is all the more interesting, because upon the body of our *Pomacanthus* there are striking colour bands, which correspond in direction very closely

Fig. 519. *Scarus* sp. Fig. 520. *Pomacanthus.*

to the lines of our new curved ordinates. In like manner, the still more bizarre outlines of other fishes of the same family of Chaetodonts will be found to correspond to very slight modifications of similar coordinates; in other words, to small variations in the values of the constants of the coaxial curves.

In Figs. 521–524 I have represented another series of Acanthopterygian fishes, not very distantly related to the foregoing. If we

XVII] THE COMPARISON OF RELATED FORMS 1063

start this series with the figure of *Polyprion*, in Fig. 521, we see that the outlines of *Pseudopriacanthus* (Fig. 522) and of *Sebastes* or *Scorpaena* (Fig. 523) are easily derived by substituting a system

Fig. 521. *Polyprion.* Fig. 522. *Pseudopriacanthus altus.*

of triangular, or radial, coordinates for the rectangular ones in which we had inscribed *Polyprion*. The very curious fish *Antigonia capros*, an oceanic relative of our own boar-fish, conforms closely to the peculiar deformation represented in Fig. 524.

Fig. 523. *Scorpaena* sp. Fig. 524. *Antigonia capros.*

Fig. 525 is a common, typical *Diodon* or porcupine-fish, and in Fig. 526 I have deformed its vertical coordinates into a system of concentric circles, and its horizontal coordinates into a system of curves which, approximately and provisionally, are made to resemble

FIGURE 6.7 – Dans un ouvrage en tous points remarquable, le lettré, biologiste et mathématicien D'Arcy Wentworth Thompson (1860-1948) souligna l'importance des facteurs environnementaux et physiques (en opposition à la seule hérédité) dans la morphogenèse des êtres vivants. La forme des poissons étant peu ou prou optimale, il n'y a pas une infinité de "plans" différents les uns des autres mais bien une poignée seulement de patrons originaux, qui permettent par des déformations (non-triviales) d'engendrer toutes les formes privilégiées par l'évolution.
Pour décrire la variabilité anatomique d'une famille ou d'une population observée, il suffit donc de donner un *template* de référence (arbitrairement complexe, mais commun à toutes les observations) et les déformations qui permettent de passer dudit template aux individus. La complexité se retrouve alors découplée en deux composantes intelligibles : une *image de référence*, complexe mais fixe ; des *déformations* propres aux sujets observées, souvent assez simples pour être décrites avec peu de paramètres.
Fait remarquable : les schémas ci-dessus présentent la variabilité des formes de poissons non pas comme des déplacements arbitraires de squelettes, mais comme **des changements de coordonnées, des déformations de l'espace *ambiant***. C'est sur cette idée mathématique de déformation *extrinsèque* de l'espace (en opposition aux mouvements intrinsèques des particules de poisson) que repose l'analyse procustéenne et la théorie "LDDMM" présentée dans ce chapitre. Grâce aux algorithmes détaillés dans les pages qui suivent, on peut aujourd'hui produire des figures de ce type de manière *automatique*.
Illustration tirée d'un livre dont je ne saurais trop vous recommander la lecture : *On Growth and Forms*, 1946.

Commençons. À tout instant t,

$$v_t \;=\; \arg\min\{\, \|v\|_k \mid \forall\, m, v(q_t^m) = v_t(q_t^m) \,\}\,. \tag{6.65}$$

Aussi, puisque v_t n'a pas de composante superflue,

$$v_t \in \{\, v \mid \forall\, m, v(q_t^m) = 0 \,\}^{\perp_k} \tag{6.66}$$

$$\text{i.e.} \qquad v_t \in \left(\bigcap_{m=1}^{M} \{\, v \mid \langle \delta_{q_t^m}, v \rangle = 0 \,\} \right)^{\perp_k}. \tag{6.67}$$

Mais on sait aussi que :

$$\langle k \star \delta_{q_t^m}, v \rangle_k \;=\; \int_{\omega \in \mathbb{R}^D} \frac{1}{\widehat{k}(\omega)}\, \overline{\widehat{k \star \delta_{q_t^m}}(\omega)} \cdot \widehat{v}(\omega)\;\, \mathrm{d}\omega \tag{6.68}$$

$$= \int_{\omega \in \mathbb{R}^D} \overline{\widehat{\delta_{q_t^m}}(\omega)} \cdot \widehat{v}(\omega)\;\, \mathrm{d}\omega \tag{6.69}$$

$$= \langle \delta_{q_t^m}, v \rangle \;=\; v(q_t^m). \tag{6.70}$$

C'est pourquoi on a, à tout instant t,

$$v_t \in \left(\bigcap_{m=1}^{M} \{\, v \mid \langle k \star \delta_{q_t^m}, v \rangle_k = 0 \,\} \right)^{\perp_k} \tag{6.71}$$

$$= \bigcup_{m=1}^{M} \left(k \star \delta_{q_t^m}\right)^{\perp_k \perp_k} \tag{6.72}$$

$$= \mathrm{Vect}\left(k \star \delta_{q_t^m},\; m \in [\![1, M]\!]\right). \tag{6.73}$$

C'est donc que l'on peut écrire

$$v_t \;=\; k \star \left(\sum_{m=1}^{M} p_t^m\, \delta_{q_t^m} \right) \;=\; k \star p_t, \tag{6.74}$$

puis

$$\|v_t\|_k^2 \;=\; \left\langle k \star p_t, k^{(-1)} \star k \star p_t \right\rangle \;=\; \langle k \star p_t, p_t \rangle \;=\; p_t^{\mathsf{T}} K_{q_t} p_t. \tag{6.75}$$

Le lien entre métrique à noyaux sur les nuages de points et métriques RKHS sur les difféomorphismes de l'espace ambiant est donc fait, d'où in fine le *Principe de Réduction*. □

Initialement développé pour décrire la dynamique des *solitons* (ou "vagues solitaires"), ce résultat lie la cométrique K_q sur les espaces de landmarks à une action de déformation de l'espace ambiant. C'est, en un sens, le **résultat fondamental** de la théorie qui fait le lien entre transport optimal et analyse procustéenne : **le cadre LDDMM, pour *Large Deformation Diffeomorphic Metric Mapping***.

Bilan Mathématique À ce point du cours, on sait maintenant penser le transport régularisé comme un problème de tir géodésique dans une variété Riemannienne de landmarks. Mais souhaite-t-on recaler parfaitement deux observations l'une sur l'autre ? En règle générale, non.

C'est que les données source et cible peuvent être corrompues, bruitées ; ou, tout simplement, que l'on ne cherche pas à modéliser la variabilité morphométrique dans ses moindres détails, afin conserver des représentations simples. Aussi, on préférera toujours se restreindre à une sous-variété $\mathcal{M}$ de l'espace de toutes les formes, ou à défaut de contrainte explicite, imposer que le *modèle* intermédiaire entre source et cible ne soit pas à une distance déraisonnable de la source.

Algorithme de recalage itératif Mais au juste, comment décomposer la variabilité entre deux formes X et Y en un chemin (la transformation) sur une variété $\mathcal{M}$ connue (l'espace des modèles, abusivement appelé "modèle") et un résiduel (assimilé à du "bruit") ? A priori, on aurait tendance à privilégier la *projection* de Y sur la variété ; munis d'une notion de distance, *d'attache aux données*, de choisir comme modèles les éléments de $\mathcal{M}$ qui sont les plus proches des observations. Autrement dit, trouver une transformation f telle que

$$X \xrightarrow{f} f(X) \leftrightarrows Y \quad \text{avec une dissimilarité minimale} \quad \text{“}\|f(X) - Y\|^2\text{”}.$$

Dans le cas des nuages de points étiquetés, on a vu au chapitre précédent que l'on pouvait choisir pour dissimilarité la distance euclidienne au carré et pour ensemble des déformations f celui des similitudes rigides : c'était le point de base de l'analyse procustéenne. Dans un cas moins favorable, celui des nuages de points sans étiquette, des courbes ou des surfaces, on pourra par exemple utiliser le formalisme des mesures, et utiliser pour dissimilarité la distance de Wasserstein donnée par le transport optimal, ou plus simplement une distance à noyau

$$\|f(X) - Y\|_s^2 \;=\; \|\mu - \nu\|_s^2 \;=\; \|B_s \star (\mu - \nu)\|^2_{L^2(\mathbb{R}^D)}\,, \tag{6.76}$$

où B_s est une fonction de flou (*blur*) de rayon caractéristique s – disons, une gaussienne. Idéalement, on sera donc à la recherche du projeté orthogonal

$$p_s^{\perp}(Y \to \mathcal{M}(X)) \;=\; \arg\min_f \|f(X) - Y\|_s^2. \tag{6.77}$$

Fonction de coût à minimiser Malheureusement, une définition des modèles comme réalisateurs de minimums contraints d'une fonction distance n'est pas très utile en pratique : au mieux, elle permet d'obtenir des équations réalisées à l'optimum (*criticité*)... que l'on ne peut résoudre dès que la dimension du problème augmente, et que $\mathcal{M}$ devient générique. C'est que pour des modèles complexes :

— La variété $\mathcal{M}$ des déformations n'est pas bien comprise de manière globale.

— On veut une certaine garantie $d_{\mathcal{M}}(X, f(X)) \;\leqslant\; C \;<\; +\infty$.

On se contentera donc de minimiser, sur la **transformation** géodésique f, la fonctionnelle d'énergie :

$$\text{Coût}(f) \;=\; \gamma_{\text{reg}} \cdot \ell^2_{\mathcal{M}}(X \to f(X)) \;+\; \gamma_{\text{att}} \cdot \|f(X) - Y\|_s^2. \tag{6.78}$$

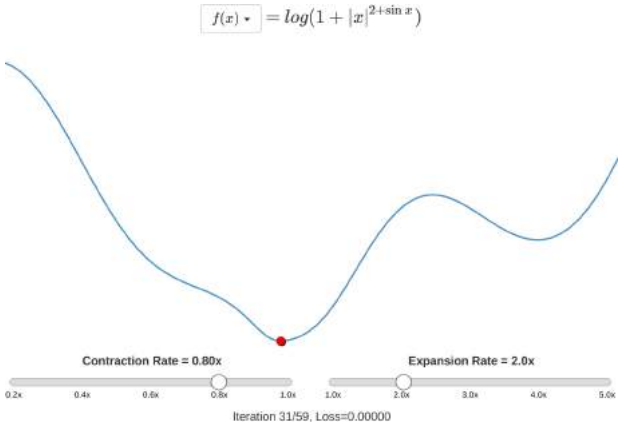
Tout l'intérêt du théorème 6.1 de structure des géodésiques est alors qu'une transformation optimale, nécessairement géodésique, est entièrement caractérisée par un **moment de tir** p_0. On peut donc réécrire le coût à minimiser sous une forme réduite, fonction du seul covecteur p_0 :

$$\text{Coût}(\mathbf{p_0}) \;=\; \gamma_{\text{reg}} \cdot p_0^{\mathsf{T}} K_{q_0} p_0 \;+\; \gamma_{\text{att}} \cdot \|\mathbf{q_1} - Y\|_s^2, \tag{6.79}$$

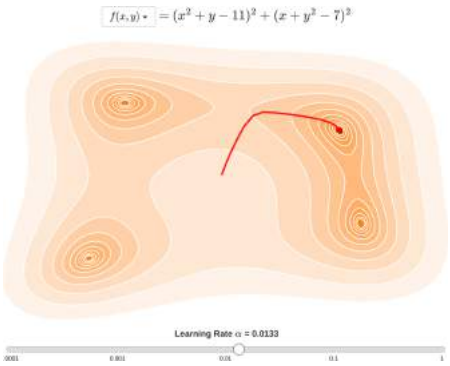
où q_1 est obtenu par intégration du flot Hamiltonien partant au temps $t = 0$ de la forme $q_0 = X$ avec un moment p_0. Si le coefficient associé au terme de régularisation est bien dominé par $\gamma_{\text{reg}} << \gamma_{\text{att}}$, le modèle *shooté* q_1 devrait être une bonne approximation du projeté défini équation (6.77).

Bilan algorithmique Grâce à une analyse mathématique forte sur la structure des trajectoires solution, le recalage fluide entre deux formes a été ramené à un simple problème de minimisation sur une variable de *moment* p_0, de *même dimension* que le nuage de points à déformer – on peut l'assimiler à un champ de "vitesses désirées".

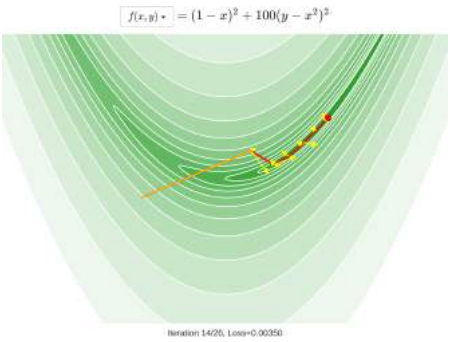
Cette minimisation sera obtenue en pratique au moyen un schéma *de descente itératif*, semblable à une descente de gradient – pour des raisons d'efficacité, on préfèrera des schémas d'ordre 2 dits de "quasi-Newton", comme L-BFGS. Pour une brève présentation de ces algorithmes, je vous conseille l'excellente page web interactive de Ben Frederickson, dont sont tirées les images ci-dessous : `www.benfrederickson.com/numerical-optimization/`. Pour une découverte des algorithmes stochastiques les plus populaires, `sebastianruder.com/optimizing-gradient-descent/` fera tout à fait l'affaire !



(a) Minimisation en dimension 1.

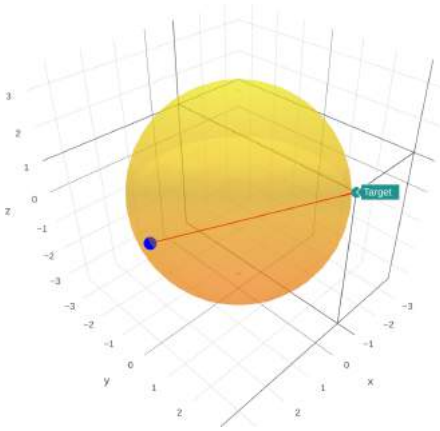


(b) Descente de gradient en dimension 2.

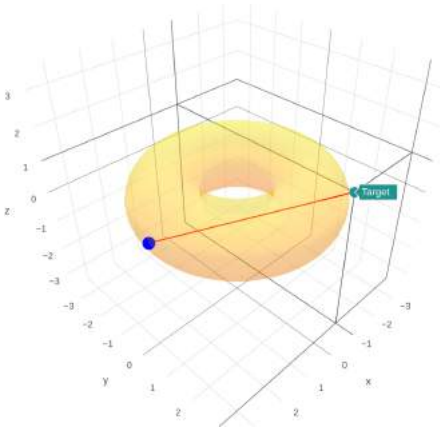


(c) Algorithme du gradient conjugué.

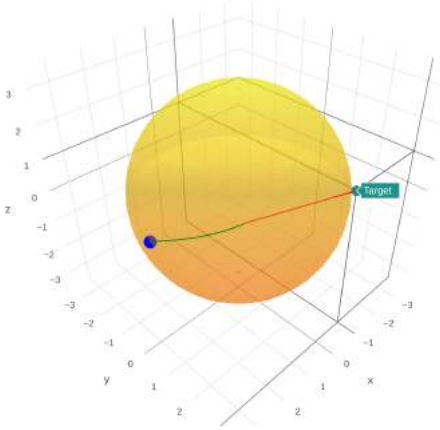
FIGURE 6.8 – Illustrations tirées de `www.benfrederickson.com/numerical-optimization/` sur les méthodes élémentaires de minimisation. (a) De manière générale, si la fonction objectif est suffisamment bien conditionné, il suffit de faire "rouler une bille" suivant la pente d'une courbe pour converger vers un minimum local. (b) Cette méthode peut être généralisée aux dimensions supérieures à moindre frais. Notez qu'en toute généralité, aucune méthode ne permettra de converger à coup sûr vers un optimum *global* : seule une optimalité *locale* pourra être garantie. (c) Pour accélérer la convergence, on pourra utiliser des méthodes plus efficace : gradient conjugué ou BFGS, par exemple.



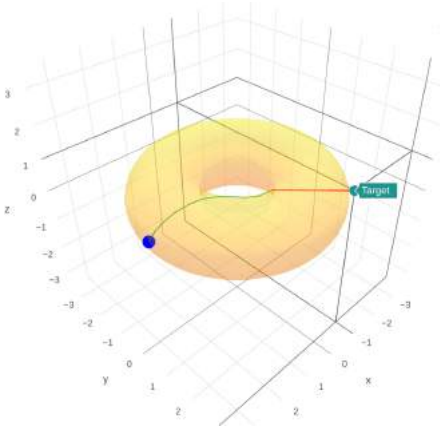
(a) Configuration initiale.



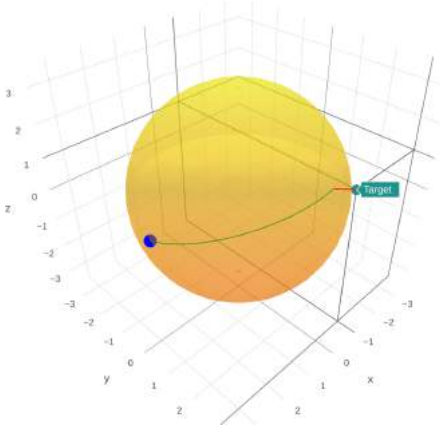
(b) Configuration initiale.



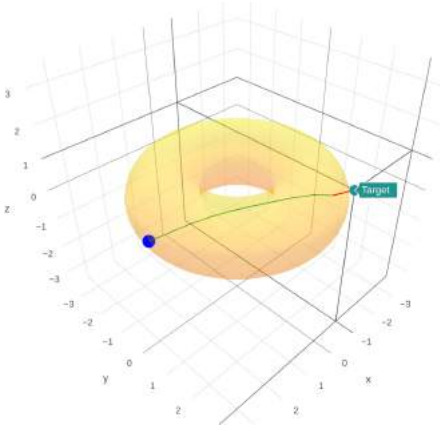
(c) Après quelques itérations.



(d) Après quelques itérations.



(e) À convergence : le quasi-projeté.



(f) À convergence : le quasi-projeté.

FIGURE 6.9 – Illustration d'une méthode de "*matching*" itérative dans un espace de dimension finie. La variété d'intérêt $\mathcal{M}$ est ici représentée comme une surface dorée : il faut la choisir pour coller au mieux à une distribution *propre* de données ; on présente ici deux exemples possibles de surface, une par colonne. La source, ou *template*, est représentée par le gros point bleu sur la variété ; la cible, ou *target*, par un point turquoise. La géodésique tirée sur la variété est en vert, et le résiduel en rouge ; à leur intersection, le *modèle*, représentant optimal de la cible dans l'orbite de la template.

Décomposition de la variabilité anatomique : l'algorithme complet

Avant d'attaquer le code proprement dit, un rappel final sur l'algorithme proposé. On note $\mathcal{M}$ l'espace de formes, orbite de q_0 sous l'action du groupe de déformations. En pratique, il s'agit de l'espace des landmarks de même cardinal que la source q_0, plus ou moins éloignés de celle-ci selon qu'ils sont facilement obtenus par une déformation k-lisse, où k est une fonction noyau.

Les variables et fonctions mises en jeu sont récapitulées dans le diagramme suivant :

$$p_0 \xrightarrow{\exp_{q_0}} q_1 \xrightarrow{\pi_c} (\mu_i, x^i) \simeq \sum_i \mu_i \delta_{x^i} \xrightarrow{W} C \in \mathbb{R} \tag{6.80}$$

Où :

— $p_0 \in T^\star_{X_0}\mathcal{M}$ est un moment de tir initial, porté par la forme X_0. En pratique, on le représentera simplement comme un vecteur $p_0 = (p_0^1, \dots, p_0^M)$ où $p_0^i \in (\mathbb{R}^2)^\star \simeq \mathbb{R}^2$ est le moment associé au i^{e} point de X_0.

— $\exp_{q_0}$ est l'exponentielle riemannienne dans $\mathcal{M}$. À un moment de tir initial p_0, elle associe l'unique point d'arrivée q_1, avec (q_t, p_t) l'unique solution du flot géodésique Hamiltonien partant de (q_0, p_0). En pratique, l'intégration du flot est implémentée comme une simple méthode d'Euler sur $(q_t, p_t) \in \mathbb{R}^{2\cdot M\times 2}$, suivant l'équation (6.41).

— Le nuage de points q_1 est donc encodé a priori comme un simple vecteur de $\mathbb{R}^{M\times 2}$.

— Une information de connectivité c, reliant les points les uns aux autres selon un graphe, est alors utilisée au travers d'un plongement π_c, qui envoie le nuage de points dans un espace de mesures approprié. Le plus simple est de procéder comme dans l'équation (6.4) : pour tout segment i encodé dans c sous la forme d'une liste d'indices (α_i, β_i), on considère le segment $[q^{\alpha_i}, q^{\beta_i}]$ et on ajoute à notre mesure un dirac positionné en son centre, de masse égale à la longueur de l'élément de forme ainsi défini. Notons que des plongements plus complexes dans des espaces de courants ou de varifolds peuvent-être utilisés pour conserver un marqueur d'*orientation* des éléments de formes.

— Enfin, on assigne un *coût* à la forme *shootée* au travers d'une fonction W, qui quantifie l'éloignement à une mesure de référence que l'on cherche à atteindre – la cible $\pi_c(Y)$. On utilise généralement le carré d'une norme duale à noyaux : une option cheap, mais localement efficace. Trouver des termes d'attache aux données pertinent selon les applications reste un sujet de recherche actif.

Notons qu'au lieu de transporter une mesure $\pi_c(q_0)$, on a ici choisi de transporter un nuage de points "sans segments", pour reconstruire la forme globale à l'arrivée via le calcul de $\pi_c(q_1)$. C'est une manière robuste et numériquement sensée d'implémenter l'action de transport "d'images" :

$$\forall\, \omega \in C_0(\mathbb{R}^d),\ \ (\varphi \cdot \mu)(\omega) = \mu\big((\omega \circ \varphi^{-1}) \cdot |\mathrm{Jac}\, \varphi^{-1}|\big)\,, \tag{6.81}$$

sans avoir à calculer de différentielle, ou d'avoir à se préoccuper qu'une courbe connexe au départ ne soit éclatée en petites écailles mal reliées à l'arrivée.

Descente de gradient pour le problème de matching Tous comptes faits, on cherche donc à minimiser sur le moment $p_0 \in (\mathbb{R}^{nd})^\star \simeq \mathbb{R}^{nd}$ la fonction de coût :

$$\text{Coût}(p_0) \ = \ \underbrace{\frac{\gamma_{\text{reg}}}{2} p_0^\intercal K_{q_0} p_0}_{\text{Régularisation}} \ + \ \underbrace{\frac{\gamma_{\text{att}}}{2} W \circ \pi_c \circ \exp_{q_0}(p_0)}_{\text{Attache aux données}}. \tag{6.82}$$

Pour cela, et en l'absence de connaissances globales sur la structure de l'énergie, on utilisera un algorithme de descente de gradient d'ordre 1 ou 2. Dans sa plus simple expression, il s'agira d'itérer sur p_0 le calcul suivant :

$$p_0(\text{it.} = n+1) \;=\; p_0(n) + \delta p_0(n), \qquad \text{avec} \;\; \delta p_0(n) = \; - \nabla^{J_{q_0}}_{p_0(n)} E, \tag{6.83}$$

où $\nabla^{J_{q_0}}_{p_0(n)} E$ est le gradient de E en $p_0(n)$, au sens d'une métrique J_{q_0} à choisir sur l'espace des moments en q_0.

Calcul du gradient On utilise la relation fondamentale de définition du gradient,

$$\forall\ \delta p_0, \quad (\nabla^J_{p_0} E)^\mathsf{T} J_{q_0} \delta p_0 = \; \mathrm{d}_{p_0} E \cdot \delta p_0 \tag{6.84}$$

$$= (\partial_{p_0} E)^\mathsf{T} \delta p_0, \tag{6.85}$$

$$\text{i.e.} \quad (\nabla^{J_{q_0}}_{p_0} E) = J^\star_{q_0}(\partial_{p_0} E). \tag{6.86}$$

S'encombrer d'une métrique J_{q_0} peut sembler un peu lourd : après tout, on a toujours l'habitude d'identifier le gradient au vecteur des dérivées partielles $\partial_{p_0} E$, représentant L^2-Riesz de la différentielle, via la métrique canonique « $J_{q_0} = \mathrm{Id}$ » – on parle alors de gradient L^2. C'est qu'ici, nous travaillons sur un espace de *moments*, sur lequel la métrique "naturelle" n'est pas l'identité, mais celle associée à la co-métrique riemannienne K_{q_0} dont on a muni l'espace de formes. On sera donc prudent et explicite dans le choix de nos notations, afin que le pas de descente $\delta p_0(n)$ reste une quantité bien comprise.

Heureusement, le vecteur des dérivées partielles du terme bilinéaire de régularisation est facile à calculer :

$$\partial_{p_0} \left[\frac{\gamma_{\mathrm{reg}}}{2} p_0^\mathsf{T} K_{q_0} p_0 \right] \;=\; \gamma_{\mathrm{reg}} K_{q_0} p_0. \tag{6.87}$$

Pour le terme d'attaches aux données, on n'a généralement pas de mal à calculer les dérivées partielles $\partial_{(\mu_i, q^i)} W = (\partial_{\mu_i} W, \partial_{q^i} W)$. Reste encore à les transporter sur p_0. On utilise pour cela la règle de dérivation d'une composée, ou *chain rule* :

$$\mathrm{d}_{p_0} \left[\frac{\gamma_{\mathrm{att}}}{2} W \circ \pi_c \circ \exp_{q_0} \right] = \frac{\gamma_{\mathrm{att}}}{2} \left(\mathrm{d}_{\pi_c \circ \exp_{q_0}(p_0)} W \right) \circ \left(\mathrm{d}_{\exp_{q_0}(p_0)} \pi_c \right) \circ \left(\mathrm{d}_{p_0} \exp_{q_0} \right), \tag{6.88}$$

soit, en passant aux adjoints/transposées :

$$\partial_{p_0} \left[\frac{\gamma_{\mathrm{att}}}{2} W \circ \pi_c \circ \exp_{q_0} \right] = \frac{\gamma_{\mathrm{att}}}{2} \left(\mathrm{d}_{p_0} \exp_{q_0} \right)^\star \circ \left(\mathrm{d}_{q_1} \pi_c \right)^\star \circ \left(\partial_{(\mu_i, q^i)} W \right). \tag{6.89}$$

Avec un peu de travail, on dispose donc de fonctions numériques explicites en la variable p_0, encodée comme un tableau réel de taille $M \times 2$:

$$p_0 \in \mathbb{R}^{M \times 2} \mapsto \mathrm{Coût}(p_0) \in \mathbb{R} \qquad \text{et} \qquad p_0 \in \mathbb{R}^{M \times 2} \mapsto \nabla_{p_0} \mathrm{Coût}(\,\cdot\,) \in \mathbb{R}^{M \times 2}, \tag{6.90}$$

que l'on pourra *plugger* dans les routines d'optimisation standard fournies par tout bon moteur de calcul scientifique.

Un véritable programme informatique

Décrire des "méthodes mathématiques" pour résoudre des problèmes concrets, c'est une chose. Les implémenter, les tester, les déboguer puis les améliorer, c'en est une autre. Pour lever tout doute sur l'applicabilité des idées présentées dans ce cours, je vous propose maintenant de voir écrites, noir sur blanc, leurs traductions sous forme de programme informatique. Les sept pages qui suivent, 500 lignes de code en tout, forment une petite *toolbox* qui nous permettra d'illustrer "clés en mains" le comportement des méthodes LDDMM aux Figures 6.10-6.12. Codé en `Python`, ce programme repose pleinement sur une fantastique bibliothèque de différenciation automatique, `theano` ; nous aurons l'occasion d'en reparler plus bas.

Fonctionnalités Pour que vous puissiez jouer sans trop de difficultés avec ce code, j'ai implémenté deux modes de chargement des données : d'abord, un classique import de fichier `.vtk` donnant les courbes sous forme de listes de points et de segments ; mais aussi, plus facile à manipuler, un mode d'importation `.png` qui extrait automatiquement les lignes de niveau d'images *bitmap* en noir et blanc.

Côté déformations, je me suis contenté du strict minimum : une cométrique à noyau isotrope, à queue lourde, avec un paramètre réglable d'échelle, `s`. N'hésitez pas modifier la formule donnée ligne 242 dans la fonction `_k(x,y,s)` !

Finalement, pour l'attache aux données ou *fidélité* qui relie le modèle `q1` à la cible `xt`, je vous propose (*brand new !*) d'utiliser une distance de type Wasserstein avec la méthode `_ot_matching`, en plus du terme "à noyau" relativement classique implémenté par la fonction `_kernel_matching`.

Plan du code Dans les pages qui suivent, tout n'est pas important loin s'en faut ! À vrai dire, les trois premières pages, incorporées au polycopié par souci d'exhaustivité, ne traitent que des problèmes d'entrée-sortie et de conversion de formats : vous pouvez les sauter. Par contre, ne manquez pas de lire les pages 4, 5 et 6 (lignes 211-406), qui transcrivent *exactement* les idées sur les variétés, les équations que nous avons détaillé dans tout le début du chapitre. Enfin, consacrée au chargement des données et à la minimisation proprement dite de l'énergie `_cost`, la septième et dernière page est d'un relatif intérêt.

Pour clarifier les différences de nature entre méthodes "`python-scipy`", routines d'affichages "`matplotlib`" et fonctions symboliques "`theano`", je me suis tenu à une convention de nommage simple : ces diverses portions de code sont respectivement intitulées sous les formes `ma_methode`, `MaRoutine` et `_ma_fonction`.

Références En quelques pages, j'ai tenté de vous fournir les bases conceptuelles de la théorie LDDMM des recalages fluides. Mais bien sûr, il manque à cet exposé les détails, les preuves que demande votre esprit de rigueur, toujours en éveil... Alors, pour me dédouaner un peu, une petite liste de références liées aux notions clés utilisées dans le code :

- D'abord, absolument essentielle, la documentation de la bibliothèque `theano`. Pour prendre en main la syntaxe, voici un lien vers le très bon tutorial officiel : `deeplearning.net/software/theano/tutorial/adding.html`.
- Ensuite, au sujet des attaches à noyau, un très bon article pratique écrit pour un public d'informaticiens : *A general framework for curve and surface comparison and registration with oriented varifolds*, CVPR2017, par Irène Kaltenmark, Benjamin Charlier et Nicolas Charon.
- Enfin, en ce qui concerne l'attache Wasserstein-Transport-Optimal, un article tout neuf : *Optimal Transport for Diffeomorphic Registration* écrit par Benjamin Charlier, F.-X. Vialard, Gabriel Peyré et moi-même ; à paraître pour la conférence MICCAI2017.

```python
# Import the relevant tools
import time                 # to measure performance
import numpy as np          # standard array library
import theano               # Autodiff & symbolic calculus library :
import theano.tensor as T   #             - mathematical tools;
from   theano import config, printing #   - printing of the Sinkhorn error.


# Display routines :
import matplotlib.cm as cm
import matplotlib.pyplot as plt
import matplotlib.colors as colors
from matplotlib.collections  import LineCollection


# Input/Output routines ========================================================================
# from '.vtk' to Curves objects  ---------------------------------------------------------------
from pyvtk import VtkData
# from '.png' to level curves  -----------------------------------------------------------------
from skimage.measure import find_contours
from scipy import misc
from scipy.ndimage.filters import gaussian_filter
from scipy.interpolate import interp1d

def arclength_param(line) :
    "Arclength parametrisation of a piecewise affine curve."
    vel = line[1:, :] - line[:-1, :]
    vel = np.sqrt(np.sum( vel ** 2, 1 ))
    return np.hstack( ( [0], np.cumsum( vel, 0 ) ) )
def arclength(line) :
    "Total length of a piecewise affine curve."
    return arclength_param(line)[-1]

def resample(line, npoints) :
    "Resamples a curve by arclength through linear interpolation."
    s = arclength_param(line)
    f = interp1d(s, line, kind = 'linear', axis = 0, assume_sorted = True)

    p = f( np.linspace(0, s[-1], npoints) )
    connec = np.vstack( (np.arange(0, len(p) - 1),
                         np.arange(1, len(p)    )) ).T
    if np.array_equal(p[0], p[-1]) : # i.e. p is a loop
        p = p[:-1]
        connec = np.vstack( (connec[:-1,:],  [len(p)-1, 0]) )
    return (p, connec)

def level_curves(fname, npoints = 200, smoothing = 10, level = 0.5) :
    "Loads regularly sampled curves from a .PNG image."
    # Find the contour lines
    img = misc.imread(fname, flatten = True) # Grayscale
    img = (img.T[:, ::-1])  / 255.
    img = gaussian_filter(img, smoothing, mode='nearest')
    lines = find_contours(img, level)

    # Compute the sampling ratio for every contour line
    lengths = np.array( [arclength(line) for line in lines] )
    points_per_line = np.ceil( npoints * lengths / np.sum(lengths) )

    # Interpolate accordingly
    points = [] ; connec = [] ; index_offset = 0
    for ppl, line in zip(points_per_line, lines) :
        (p, c) = resample(line, ppl)
        points.append(p)
        connec.append(c + index_offset)
        index_offset += len(p)

    size   = np.maximum(img.shape[0], img.shape[1])
    points = np.vstack(points) / size
    connec = np.vstack(connec)
    return Curve(points, connec)
```

```
# Pyplot Output ================================================================================

def GridData() :
    "Returns the coordinates and connectivity of the grid carried along by a deformation."
    nlines = 11 ; ranges = [ (0,1), (0,1) ] # one square = (.1,.1)
    np_per_lines = (nlines-1) * 4 + 1       # Supsample lines to get smooth figures
    x_l = [np.linspace(min_r, max_r, nlines      ) for (min_r,max_r) in ranges]
    x_d = [np.linspace(min_r, max_r, np_per_lines) for (min_r,max_r) in ranges]

    v = [] ; c = [] ; i = 0
    for x in x_l[0] :                   # One vertical line per x :
        v += [ [x, y] for y in x_d[1] ]  # Add points to the list of vertices.
        c += [ [i+j,i+j+1] for j in range(np_per_lines-1)] # + appropriate connectivity
        i += np_per_lines
    for y in x_l[1] :                   # One horizontal line per y :
        v += [ [x, y] for x in x_d[1] ]  # Add points to the list of vertices.
        c += [ [i+j,i+j+1] for j in range(np_per_lines-1)] # + appropriate connectivity
        i += np_per_lines

    return ( np.vstack(v), np.vstack(c) ) # (vertices, connectivity)

def ShowTransport( Q, Xt, Gamma, ax ) :
    "Displays a transport plan."
    points = [] ; connectivity = [] ; curr_id = 0
    Q_points,Q_weights = Q.to_measure()  ;  xtpoints = Xt.points # Extract the centers + areas
    for (a, mui, gi) in zip(Q_points, Q_weights, Gamma) :
        gi = gi / mui # gi[j] = fraction of the mass from "a" which goes to xtpoints[j]
        for (seg, gij) in zip(Xt.connectivity, gi) :
            mass_per_line = 0.05
            if gij >= mass_per_line :
                nlines = np.floor(gij / mass_per_line)
                ts     = np.linspace(.35, .65, nlines)
                for t in ts :
                    b = (1-t) * xtpoints[seg[0]] + t * xtpoints[seg[1]]
                    points += [a, b]; connectivity += [[curr_id, curr_id + 1]]; curr_id += 2
    if len(connectivity) > 0 :
        Plan = Curve(np.vstack(points), np.vstack(connectivity))
        Plan.plot(ax, color = (.6,.8,1.), linewidth = 1)

def DisplayShoot(Q0, G0, p0, Q1, G1, Xt, info, it, scale_momentum, scale_attach) :
    "Displays a pyplot Figure and save it."
    # Figure at "t = 0" : ----------------------------------------------------------------------
    fig = plt.figure(1, figsize = (10,10), dpi=100); fig.clf(); ax = fig.add_subplot(1, 1, 1)
    ax.autoscale(tight=True)

    G0.plot(ax, color = (.8,.8,.8), linewidth = 1)
    Xt.plot(ax, color = (.85, .6, 1.))
    Q0.plot(ax)
    ax.quiver( Q0.points[:,0], Q0.points[:,1], p0[:,0], p0[:,1],
               scale = scale_momentum, color='blue')

    ax.axis([0, 1, 0, 1]) ; ax.set_aspect('equal') ; plt.draw() ; plt.pause(0.001)
    fig.savefig( 'output/momentum_' + str(it) + '.png' )

    # Figure at "t = 1" : ----------------------------------------------------------------------
    fig = plt.figure(2, figsize = (10,10), dpi=100); fig.clf(); ax = fig.add_subplot(1, 1, 1)
    ax.autoscale(tight=True)

    if scale_attach == 0 : # Convenient way of saying that we're using a transport plan.
        ShowTransport( Q1, Xt, info, ax)
    else :                 # Otherwise, it's a kernel matching term.
        ax.imshow(info, interpolation='bilinear', origin='lower',
                  vmin = -scale_attach, vmax = scale_attach, cmap=cm.RdBu,
                  extent=(0,1, 0, 1))
    G1.plot(ax, color = (.8,.8,.8), linewidth = 1)
    Xt.plot(ax, color = (.76, .29, 1.))
    Q1.plot(ax)

    ax.axis([0, 1, 0, 1]) ; ax.set_aspect('equal') ; plt.draw() ; plt.pause(0.001)
    fig.savefig( 'output/model_' + str(it) + '.png' )
```

```
# Curve representations ================================================================================

class Curve :
    "Encodes a 2D curve as an array of float coordinates + a connectivity list."
    def __init__(self, points, connectivity) :
        "points should be a n-by-2 float array, connectivity an nsegments-by-2 int array."
        self.points       = points
        self.connectivity = connectivity

    def segments(self) :
        "Returns the list of segments the curve is made of."
        return np.array( [  [self.points[l[0]], self.points[l[1]]] for l in self.connectivity ] )

    def to_measure(self) :
        """
        Outputs the sum-of-diracs measure associated to the curve.
        Each segment from the connectivity matrix self.c
        is represented as a weighted dirac located at its center,
        with weight equal to the segment length.
        """
        segments = self.segments()
        centers = [         .5 * (  seg[0] + seg[1]      ) for seg in segments ]
        lengths = [np.sqrt(np.sum( (seg[1] - seg[0])**2 )) for seg in segments ]
        return ( np.array(centers), np.array(lengths) )

    @staticmethod
    def _vertices_to_measure( q, connec ) :
        """
        Transforms a theano array 'q1' into a measure, assuming a connectivity matrix connec.
        It is the Theano equivalent of 'to_measure' : as theano only handles numeric arrays,
        it could not be implemented in a neat Object-Oriented fashion.
        """
        a = q[connec[:,0]] ; b = q[connec[:,1]]
        # A curve is represented as a sum of diracs, one for each segment
        x  = .5 * (a + b)                 # Mean
        mu = T.sqrt( ((b-a)**2).sum(1) )  # Length
        return (x, mu)

    def plot(self, ax, color = 'rainbow', linewidth = 3) :
        "Simple display using a per-id color scheme."
        segs = self.segments()

        if color == 'rainbow' :   # rainbow color scheme to see pointwise displacements
            ncycles    = 5
            cNorm      = colors.Normalize(vmin=0, vmax=(len(segs)-1)/ncycles)
            scalarMap  = cm.ScalarMappable(norm=cNorm, cmap=plt.get_cmap('hsv') )
            seg_colors = [ scalarMap.to_rgba( i % ((len(segs)-1)/ncycles) )
                           for i in range(len(segs)) ]
        else :                    # uniform color
            seg_colors = [ color for i in range(len(segs)) ]

        line_segments = LineCollection(segs, linewidths=(linewidth,),
                                       colors=seg_colors, linestyle='solid')
        ax.add_collection(line_segments)

    @staticmethod
    def from_file(fname) :
        if   fname[-4:] == '.png' :
            return level_curves(fname)
        elif fname[-4:] == '.vtk' :
            data = VtkData(fname)
            points = np.array(data.structure.points)[:,0:2] # Discard "Z"
            connec = np.array(data.structure.polygons)
            return Curve((points + 150)/300, connec)
```

```
# My .theanorc reads as follow :
# [nvcc]
# flags=-D_FORCE_INLINES
#
# [global]
# device=cuda
# floatX=float32
#
# The first section, copy-pasted from the "easy-install on Ubuntu", is supposed to fix a gcc bug.
# The second one allows me to use my GPU as the default computing device in float32 precision.
#
# On my Dell laptop, it is a GeForce GTX 960M with 640 Cuda cores and 2Gb of memory.


# Theano is a fantastic deep learning library : it transforms symbolic python code
# into highly optimized CPU/GPU binaries, which are called in the background seamlessly.
#
# We now show how to code a whole LDDMM pipeline into three pages of theano symbolic code.

# Part 1 : cometric on the space of landmarks, kinetic energy on the phase space (Hamiltonian)===


def _squared_distances(x, y) :
    "Returns the matrix of |x_i-y_j|^2."
    x_col = x.dimshuffle(0, 'x', 1)
    y_lin = y.dimshuffle('x', 0, 1)
    return T.sum( (x_col - y_lin)**2 , 2 )

def _k(x, y, s) :
    "Returns the matrix of k(x_i,y_j)= 1/(1+|x_i-y_j|^2)^{1/4}, with a heavy tail."
    sq = _squared_distances(x, y) / (s**2)
    return T.pow( 1. / ( 1. + sq ), .25 )

def _cross_kernels(q, x, s) :
    "Returns the full k-correlation matrices between two point clouds q and x."
    K_qq = _k(q, q, s)
    K_qx = _k(q, x, s)
    K_xx = _k(x, x, s)
    return (K_qq, K_qx, K_xx)

def _Hqp(q, p, sigma) :
    "The hamiltonian, or kinetic energy of the shape q with momenta p."
    pKqp =  _k(q, q, sigma) * (p.dot(p.T))# Use a simple isotropic kernel
    return .5 * T.sum(pKqp)               # H(q,p) = 1/2 · sum_{i,j} k(x_i,x_j)p_i.p_j


# Part 2 : Geodesic shooting ========================================================================
# The partial derivatives of the Hamiltonian are automatically computed !
def _dq_Hqp(q,p,sigma) :
    return T.grad(_Hqp(q,p,sigma), q)
def _dp_Hqp(q,p,sigma) :
    return T.grad(_Hqp(q,p,sigma), p)

def _hamiltonian_step(q,p, sigma) :
    "Simplistic euler scheme step with dt = .1."
    return [q + .1 * _dp_Hqp(q,p,sigma) ,  # See eq. (6.41)
            p - .1 * _dq_Hqp(q,p,sigma) ]

def _HamiltonianShooting(q, p, sigma) :
    "Shoots to time 1 a k-geodesic starting (at time 0) from q with momentum p."
    # Here, we use the "scan" theano routine, which  can be understood as a "for" loop
    result, updates = theano.scan(fn            = _hamiltonian_step,
                                  outputs_info  = [q,p],
                                  non_sequences = sigma,
                                  n_steps       = 10            )  # We hardcode the "dt = .1"
    final_result = [result[0][-1], result[1][-1]]  # We do not store the intermediate results
    return final_result                            # and only return the final state + momentum
```

```
# Part 2bis : Geodesic shooting + deformation of the ambient space, for visualization ===========
def _HamiltonianCarrying(q0, p0, grid0, sigma) :
    """
    Similar to _HamiltonianShooting, but also conveys information about the deformation of
    an arbitrary point cloud 'grid' in the ambient space.
    """
    def _carrying_step(q,p,g,s) :
        "Simplistic euler scheme step with dt = .1."
        return [q + .1 * _dp_Hqp(q,p, s),  p - .1 * _dq_Hqp(q,p, s), g + .1 * _k(g, q, s).dot(p)]
    # Here, we use the "scan" theano routine, which  can be understood as a "for" loop
    result, updates = theano.scan(fn            = _carrying_step,
                                  outputs_info  = [q0,p0,grid0],
                                  non_sequences = sigma,
                                  n_steps       = 10           ) # We hardcode the "dt = .1"
    final_result = [result[0][-1], result[1][-1], result[2][-1]] # Don't store intermediate steps
    return final_result                             # return the final state + momentum + grid

# Part 3 : Data attachment ========================================================================

def _ot_matching(q1_x, q1_mu, xt_x, xt_mu, radius) :
    """
    Given two measures q1 and xt represented by locations/weights arrays,
    outputs an optimal transport fidelity term and the transport plan.
    """
    # The Sinkhorn algorithm takes as input three Theano variables :
    c = _squared_distances(q1_x, xt_x) # Wasserstein cost function
    mu = q1_mu ; nu = xt_mu

    # Parameters of the Sinkhorn algorithm.
    epsilon            = (.02)**2          # regularization parameter
    rho                = (.5) **2          # unbalanced transport (See PhD Th. of Lenaic Chizat)
    niter              = 10000             # max niter in the sinkhorn loop
    tau                = -.8               # Nesterov-like acceleration

    lam = rho / (rho + epsilon)            # Update exponent

    # Elementary operations .....................................................................
    def ave(u,u1) :
        "Barycenter subroutine, used by kinetic acceleration through extrapolation."
        return tau * u + (1-tau) * u1
    def M(u,v)  :
        "M_{ij} = (-c_{ij} + u_i + v_j) / \epsilon"
        return (-c + u.dimshuffle(0,'x') + v.dimshuffle('x',0)) / epsilon
    lse = lambda A    : T.log(T.sum( T.exp(A), axis=1 ) + 1e-6) # slight modif to prevent NaN

    # Actual Sinkhorn loop ......................................................................
    # Iteration step :
    def sinkhorn_step(u, v, foo) :
        u1=u # useful to check the update
        u = ave( u, lam * ( epsilon * ( T.log(mu) - lse(M(u,v))    ) + u ) )
        v = ave( v, lam * ( epsilon * ( T.log(nu) - lse(M(u,v).T) ) + v ) )
        err = T.sum(abs(u - u1))

        return (u,v,err), theano.scan_module.until(err < 1e-4) # "break" the loop if error < tol

    # Scan = "For loop" :
    err0 = np.arange(1, dtype=config.floatX)[0]
    result, updates = theano.scan( fn            = sinkhorn_step,            # Iterated routine
                                   outputs_info  = [(0.*mu), (0.*nu), err0], # Starting estimates
                                   n_steps       = niter                   # Number of iterations
                                 )
    U, V = result[0][-1], result[1][-1] # We only keep the final dual variables
    Gamma = T.exp( M(U,V) )              # Eventual transport plan g = diag(a)*K*diag(b)
    cost  = T.sum( Gamma * c )          # Simplistic cost, chosen for readability in this tutorial
    if True :
        print_err_shape = printing.Print('error  : ', attrs=['shape'])
        errors          = print_err_shape(result[2])
        print_err  = printing.Print('error  : ') ; err_fin  = print_err(errors[-1])
        cost += .00000001 * err_fin   # hack to prevent the pruning of the error-printing node...
    return [cost, Gamma]
```

```
def _kernel_matching(q1_x, q1_mu, xt_x, xt_mu, radius) :
    """
    Given two measures q1 and xt represented by locations/weights arrays,
    outputs a kernel-fidelity term and an empty 'info' array.
    """
    K_qq, K_qx, K_xx = _cross_kernels(q1_x, xt_x, radius)
    q1_mu = q1_mu.dimshuffle(0,'x')  # column
    xt_mu = xt_mu.dimshuffle(0,'x')  # column
    cost = .5 * (   T.sum(K_qq * q1_mu.dot(q1_mu.T)) \
                  +  T.sum(K_xx * xt_mu.dot(xt_mu.T)) \
                  -2*T.sum(K_qx * q1_mu.dot(xt_mu.T))  )

    # Info = the 2D graph of the blurred distance function
    res    = 100 ; ticks = np.linspace( 0, 1, res + 1)[:-1] + 1/(2*res)
    X,Y    = np.meshgrid( ticks, ticks )
    points = T.TensorConstant( T.TensorType( config.floatX, [False,False] ) ,
                               np.vstack( (X.ravel(), Y.ravel()) ).T.astype(config.floatX) )

    info   = _k( points, q1_x , radius ).dot(q1_mu) \
           - _k( points, xt_x , radius ).dot(xt_mu)
    return [cost , info.reshape( (res,res) ) ]

def _data_attachment(q1_measure, xt_measure, radius) :
    "Given two measures and a radius, returns a cost - as a Theano symbolic variable."
    if radius == 0 : # Convenient way to allow the choice of a method
        return _ot_matching(q1_measure[0], q1_measure[1],
                            xt_measure[0], xt_measure[1],
                            radius)
    else :
        return _kernel_matching(q1_measure[0], q1_measure[1],
                            xt_measure[0], xt_measure[1],
                            radius)

# Part 4 : Cost function and derivatives ============================================================

def _cost( q,p, xt_measure, connec, params ) :
    """
    Returns a total cost, sum of a small regularization term and the data attachment.
    .. math ::

        C(q_0, p_0) = .01 * H(q0,p0) + 1 * A(q_1, x_t)

    Needless to say, the weights can be tuned according to the signal-to-noise ratio.
    """
    s,r  = params                          # Deformation scale, Attachment scale
    q1 = _HamiltonianShooting(q,p,s)[0]  # Geodesic shooting from q0 to q1
    # To compute a data attachment cost, we need the set of vertices 'q1' into a measure.
    q1_measure  = Curve._vertices_to_measure( q1, connec )
    attach_info = _data_attachment( q1_measure,  xt_measure,  r )
    return [ .01* _Hqp(q, p, s) + 1* attach_info[0] , attach_info[1] ] # [cost, info]

# The discrete backward scheme is automatically computed :
def _dcost_p( q,p, xt_measure, connec, params ) :
    "The gradients of C wrt. p_0 is automatically computed."
    return T.grad( _cost(q,p, xt_measure, connec, params)[0] , p)

#=================================================================================================

def VisualizationRoutine(Q0, params) :
    print('Compiling the ShootingVisualization routine.')
    time1 = time.time()
    q, p, grid = T.matrices('q', 'p', 'g') #  assign types to the teano variables
    ShootingVisualization = theano.function([q,p, grid],                                  # input
                                            _HamiltonianCarrying(q, p, grid, params[0]),  # output
                                            allow_input_downcast=True)  # GPU = float32 only,
                                    # whereas numpy uses float64 : we allow silent conversion
    time2 = time.time()
    print('Compiled in : ', '{0:.2f}'.format(time2 - time1), 's')
    return ShootingVisualization

```

```
def perform_matching( Q0, Xt, params, scale_momentum = 1, scale_attach = 1) :
    "Performs a matching from the source Q0 to the target Xt, returns the optimal momentum P0."
    (Xt_x, Xt_mu) = Xt.to_measure()      # Transform the target into a measure once and for all
    q0 = Q0.points ; p0 = np.zeros(q0.shape)    # Null initialization for the shooting momentum

    # Compilation ----------------------------------------------------------------------------
    print('Compiling the energy functional.')
    time1 = time.time()
    # Cost is a function of 6 parameters :
    # The source 'q',                    the starting momentum 'p',
    # the target points 'xt_x',          the target weights 'xt_mu',
    # the deformation scale 'sigma_def', the attachment scale 'sigma_att'.
    q, p, xt_x = T.matrices('q', 'p', 'xt_x') ; xt_mu = T.vector('xt_mu') #  assign types

    # Compilation. Depending on settings specified in the ~/.theanorc file or explicitely given
    # at execution time, this will produce CPU or GPU code under the hood.
    Cost  = theano.function([q,p, xt_x,xt_mu ],
                            [   _cost( q,p, (xt_x,xt_mu), Q0.connectivity, params )[0],
                             _dcost_p( q,p, (xt_x,xt_mu), Q0.connectivity, params )   ,
                                _cost( q,p, (xt_x,xt_mu), Q0.connectivity, params )[1] ],
                            allow_input_downcast=True)
    time2 = time.time()
    print('Compiled in : ', '{0:.2f}'.format(time2 - time1), 's')

    # Display pre-computing ------------------------------------------------------------------
    connec = Q0.connectivity ; q0 = Q0.points ; g0,cgrid = GridData() ; G0 = Curve(g0, cgrid )
    # Given q0, p0 and grid points grid0 , outputs (q1,p1,grid1) after the flow
    # of the geodesic equations from t=0 to t=1 :
    ShootingVisualization = VisualizationRoutine(q0, params)

    # L-BFGS minimization --------------------------------------------------------------------
    from scipy.optimize import minimize
    def matching_problem(p0_vec) :
        "Energy minimized in the variable 'p0'."
        p0 = p0_vec.reshape(q0.shape)
        [c, dp_c, info] = Cost(q0, p0, Xt_x, Xt_mu)
        matching_problem.Info = info

        if (matching_problem.it % 1 == 0) and (c < matching_problem.bestc) :
            matching_problem.bestc = c
            q1,p1,g1 = ShootingVisualization(q0, p0, np.array(g0))
            Q1 = Curve(q1, connec) ; G1 = Curve(g1, cgrid )
            DisplayShoot( Q0, G0, p0, Q1, G1, Xt, info,
                          matching_problem.it, scale_momentum, scale_attach)

        print('Iteration : ', matching_problem.it, ', cost : ', c, ' info : ', info.shape)
        matching_problem.it += 1
        # The fortran routines used by scipy.optimize expect float64 vectors
        # instead of the gpu-friendly float32 matrices, so we need a slight conversion
        return (c, dp_c.ravel().astype('float64'))
    matching_problem.bestc = np.inf ; matching_problem.it = 0 ; matching_problem.Info = None

    time1 = time.time()
    res = minimize( matching_problem,     # function to minimize
                    p0.ravel(),           # starting estimate
                    method = 'L-BFGS-B',  # an order 2 method
                    jac = True,           # matching_problems also returns the gradient
                    options = dict(
                        maxiter = 1000,   # max number of iterations
                        ftol    = .000001,# Don't bother fitting the shapes to float precision
                        maxcor  = 10      # Number of previous grads used to approx. the Hessian
                    ))
    time2 = time.time()

    p0 = res.x.reshape(q0.shape)
    print('Convergence success  : ', res.success, ', status = ', res.status)
    print('Optimization message : ', res.message.decode('UTF-8'))
    print('Final cost   after ', res.nit, ' iterations : ', res.fun)
    print('Elapsed time after ', res.nit, ' iterations : ', '{0:.2f}'.format(time2 - time1), 's')
    return p0, matching_problem.Info
```

```
491 def matching_demo(source_file, target_file, params, scale_mom = 1, scale_att = 1) :
492     Q0 = Curve.from_file(source_file) # Load source...
493     Xt = Curve.from_file(target_file) # and target.
494 
495     # Compute the optimal shooting momentum :
496     p0, info = perform_matching( Q0, Xt, params, scale_mom, scale_att)
497 
498 if __name__ == '__main__' :
499     plt.ion()
500     plt.show()
501     matching_demo('australopithecus.vtk','sapiens.vtk', (.05,.01), scale_mom = .3,scale_att = .1)
502     matching_demo('amoeba_1.png',        'amoeba_2.png',(.05,  0), scale_mom = .3, scale_att = 0)
```

Bilan : donnez une chance à Theano ! Que retenir de ces sept pages de code ? D'abord, que les mathématiciens disposent aujourd'hui d'outils formidablement adaptés, ici la combinaison du puissant langage de script `python` et de la bibliothèque de calcul symbolique `theano`. Diablement efficace, cette dernière a "changé ma vie" – si, si...

Voyez donc ! Jusqu'au début des années 2010, pour implémenter un algorithme de ce type de manière vraiment efficace, un chercheur était obligé d'écrire en détail la formule du coût à minimiser en fonction des données... Mais aussi de calculer *à la main* ses premières dérivées (souvent horribles), et de consacrer des pages et des pages de code à leurs implémentations. À déboguer, un enfer.

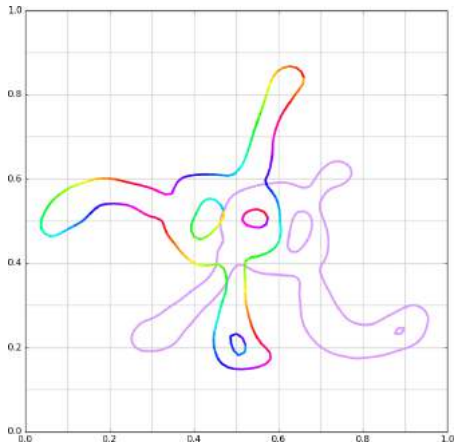
Aujourd'hui, ce travail pénible (vous pouvez me croire là-dessus...) a été remplacé par une seule ligne de magie, `theano.function(...)`. Lançant en sous-main un moteur d'optimisation avancé, elle va faire appel à un puissant algorithme ad hoc pour optimiser le graphe de calcul, avant de compiler le tout en code machine via `gcc` ; résultat : une routine numérique précise et diablement efficace. Pour tirer parti des cartes graphiques NVidia massivement parallèles, il n'y a même plus vraiment besoin d'apprendre à coder en CUDA : la bibliothèque s'en charge pour vous.

Tout cela résulte d'une évolution bien naturelle : à mesure qu'un domaine mûrit, que les enjeux industriels émergent et incitent de nombreux ingénieurs à travailler sur le sujet, des outils "métiers" de qualité sont développés. Dans notre cas, l'émergence d'outils de différentiation automatique performants est liée à l'intérêt suscité par les applications du *Deep Learning*. Dans d'autres branches des mathématiques, on peut citer les outils de vérification automatique de preuve et de simulation de fluides qui sont poussés par les demandes de l'industrie aéronautique (pilotes automatiques, soufflerie) : nous avions présenté le logiciel *Solidworks* dans la Figure 10.11.

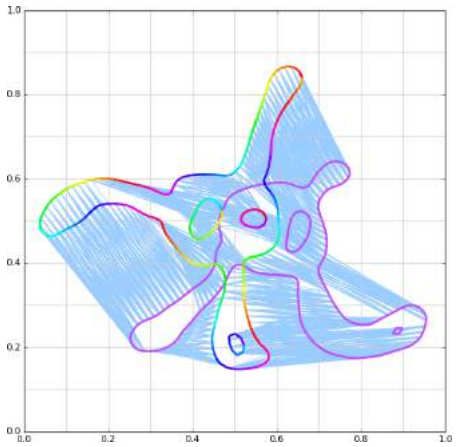
L'essentiel est que nous puissions maintenant nous concentrer sur notre *cœur de métier* : les modèles, les mathématiques. Une nouvelle idée ? Aussitôt dit, aussitôt fait : le numérique est devenu aussi malléable que la craie.

Quelques résultats Les aspects "computationnels" étant maintenant bien compris, sous contrôle, il s'agit pour nous d'explorer l'espace des paramètres ; de décrire l'influence de la fonction de noyau `_k(·,·,s)`, de la fidélité `_data_attachment` et du schéma d'optimisation `minimize` sur les recalages obtenus. Plus généralement, d'interroger les capacités du modèle LDDMM de déformations *riemanniennes* de l'espace ambiant. C'est, vous vous en doutez, le sujet de nombreux articles : dans les pages qui suivent, on se contentera de mettre en évidence le rôle du premier paramètre de la théorie, le rayon caractéristique "$\sigma =$ `s`" de la fonction de noyau k associée aux déformations.

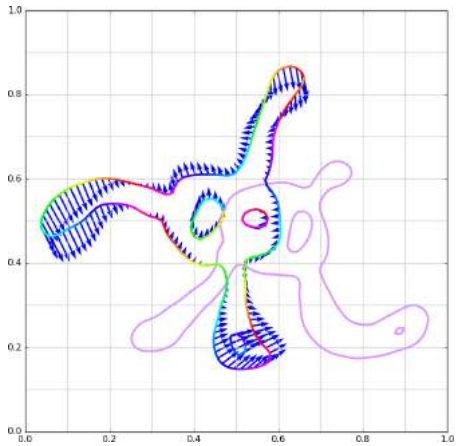
```
matching_demo('amoeba_1.png', 'amoeba_2.png', (.05, 0), scale_mom = .3, scale_att = 0)
```



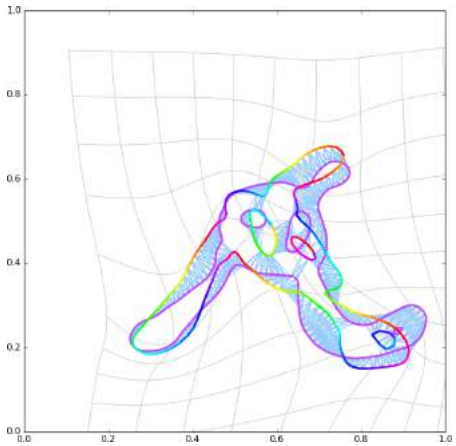
(a) Itération 0, configuration initiale.



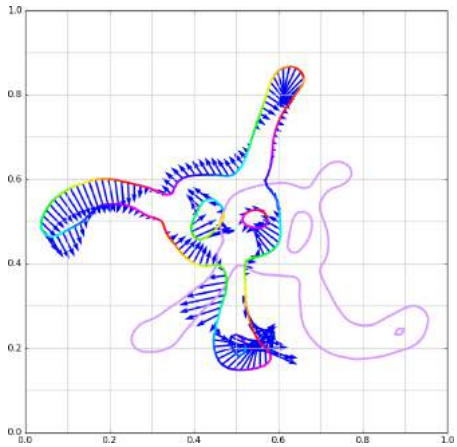
(b) Itération 0, attache aux données.



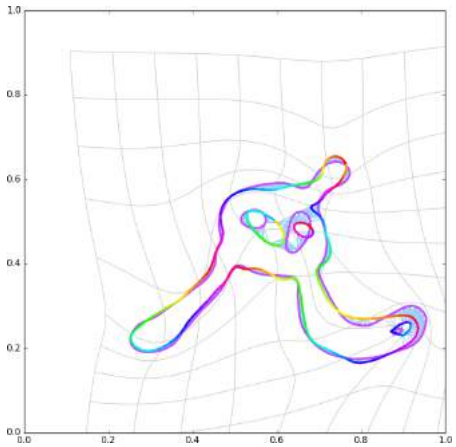
(c) Itération 15, moment de tir.



(d) Itération 15, modèle en recalage.



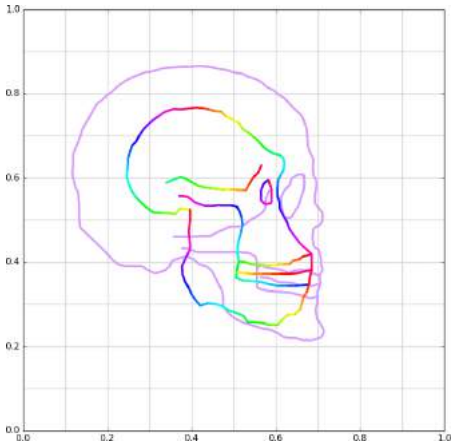
(e) Itération 80, moment de tir.



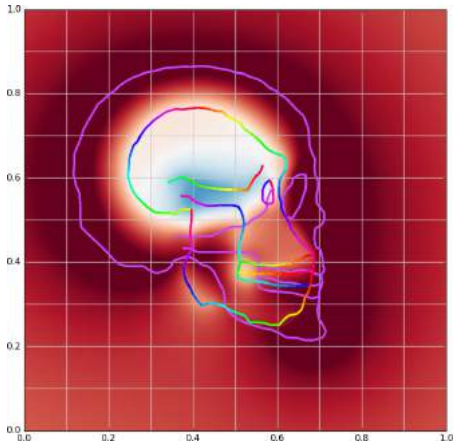
(f) Itération 80, modèle recalé.

FIGURE 6.10 – Recalage de deux silhouettes synthétiques d'amibes. On utilise ici un noyau de déformation k de rayon caractéristique $\sigma =$ `.05`, avec une attache aux données de type Wasserstein-Sinkhorn (transport optimal régularisé). Représentés en bleu ciel, les plans de transport calculés entre les modèles `q1` et la cible `xt` agissent comme des ressorts pour *tirer* le modèle, guider le moment de tir `p0` au cours de la minimisation. Pour des questions d'efficacité algorithmique, on se contente ici de calculer des plans de transport *diffus*. Le recalage final n'est donc pas très précis mais capture bien les grandes déformations des bras : en pratique, on peut utiliser cette méthode comme un pré-recalage souple et robuste.

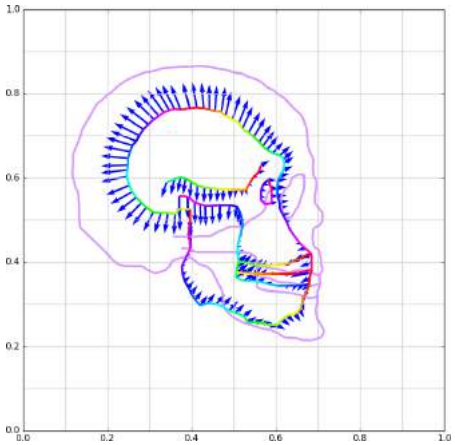
```
# Matching Skulls :
matching_demo('australopithecus.vtk','sapiens.vtk', (.05,.01), scale_mom = .3,scale_att = .1)
```



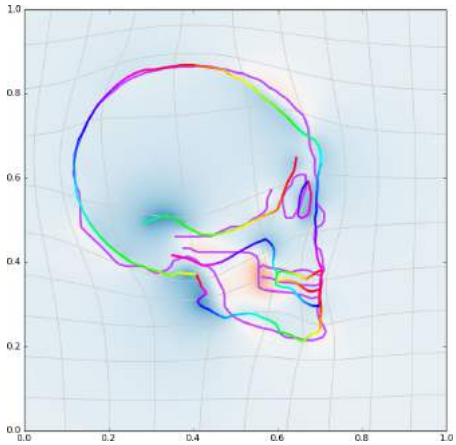
(a) Itération 0, configuration initiale.



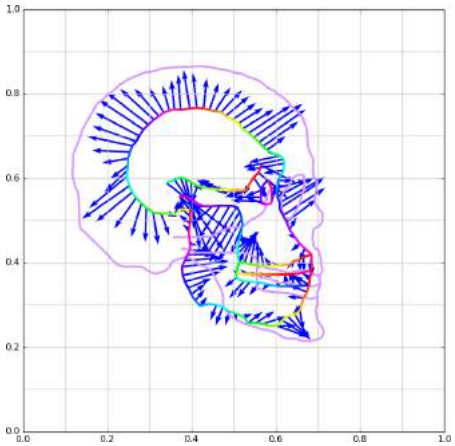
(b) Itération 0, attache aux données.



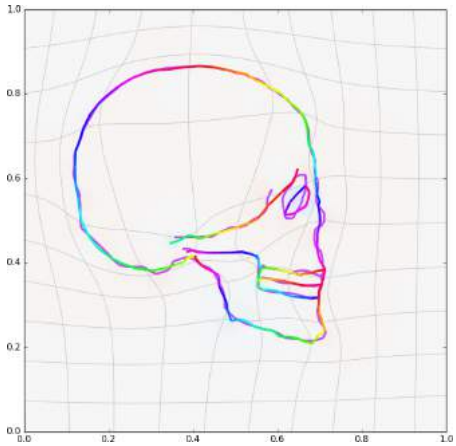
(c) Itération 25, moment de tir.



(d) Itération 25, modèle en recalage.



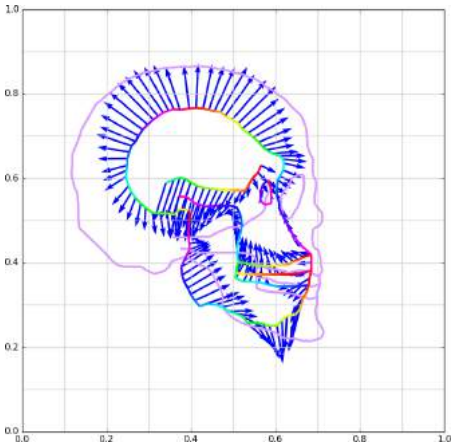
(e) Itération 200, moment de tir.



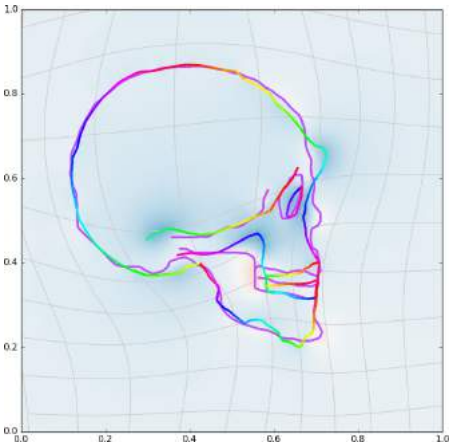
(f) Itération 200, modèle recalé.

FIGURE 6.11 – Recalage d'un crâne d'australopithèque sur un homo sapiens. On utilise ici un noyau de déformation k de rayon $\sigma =$ `.05`, avec une attache à noyau de rayon `.01`. L'attache aux données est figurée par la différence des aires d'influence : le rouge et le bleu figurent des excès de masse pour la cible `xt` et le modèle `q1`, que l'algorithme essaie de recaler au mieux.
Données gracieusement fournies par l'équipe Aramis, Institut du Cerveau et de la Moelle épinière.

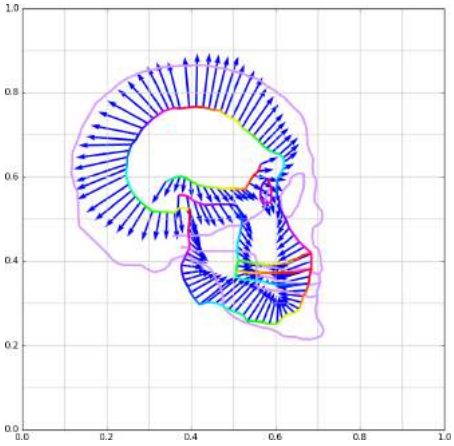
```
matching_demo('australopithecus.vtk','sapiens.vtk', (.25,.01), scale_mom = .6,scale_att = .1)
matching_demo('australopithecus.vtk','sapiens.vtk', ( .5,.01), scale_mom =1.5,scale_att = .1)
matching_demo('australopithecus.vtk','sapiens.vtk', ( 1.,.01), scale_mom =1.5,scale_att = .1)
```



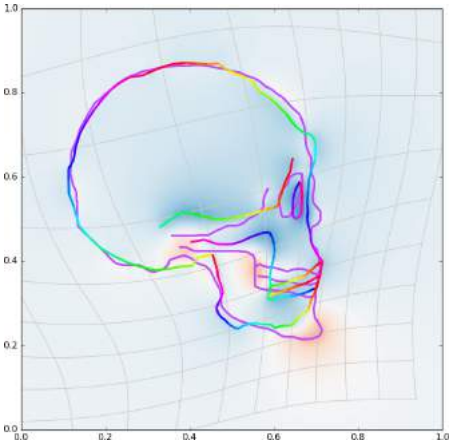
(a) $\sigma =$ `.25`, moment de tir.



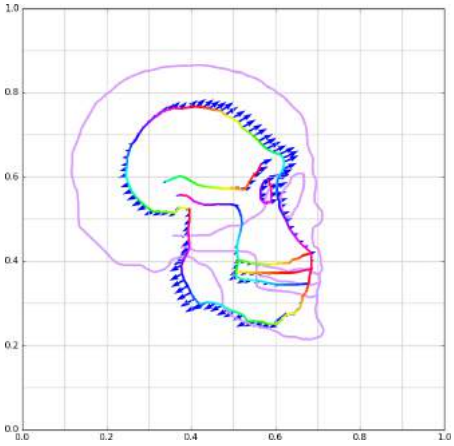
(b) $\sigma =$ `.25`, modèle recalé.



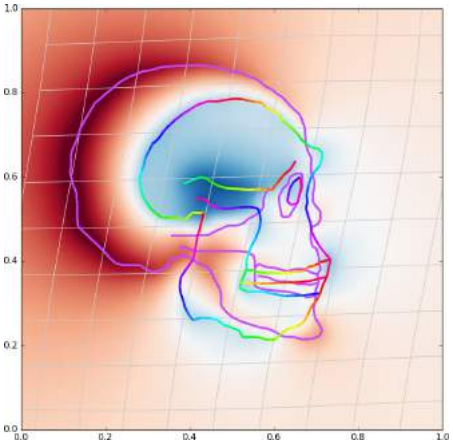
(c) $\sigma =$ `.5`, moment de tir.



(d) $\sigma =$ `.5`, modèle recalé.



(e) $\sigma =$ `1.`, moment de tir.



(f) $\sigma =$ `1.`, modèle recalé.

FIGURE 6.12 – Recalages finaux avec la même attache aux données, mais des noyaux de déformation de plus en plus gros. En rigidifiant toujours plus nos champs de vecteurs, on restreint l'espace des déformations pour finalement retrouver l'ensemble des *translations* rigides. En ce sens, la théorie exposée ici est bien une généralisation de l'analyse *procustéenne* développée au chapitre 5.

Applications en imagerie médicale

Nous avons vu qu'il est possible de munir un espace de formes d'une métrique, d'une distance qui pénalise les dissimilarités géométriques. Étant donnés deux formes "qui se ressemblent", on peut *calculer* un chemin *géodésique* optimal qui minimise un critère de régularité bien choisi, et considérer que la distance entre celles-ci est précisément donné par le coût – ou *longueur* – de la *déformation* plus un résiduel en dehors du modèle, traité comme du *bruit*. Mais à quoi bon arpenter ainsi des espaces de formes génériques ?

Régression et suivi longitudinal On connaît tous les courbes de croissances présentes sur nos carnets de santé : dûment remplies, elles permettent de suivre notre évolution, de la comparer à une population "normale" et de détecter un éventuel retard de croissance – voir Figure 6.14. Maintenant, serait-il possible de faire la même chose avec, disons, la forme du cœur ? Attention : il ne s'agit pas ici de remplacer une grandeur scalaire – la taille – par un simple vecteur "largeur/longueur/volume/que-sais-je". Ce traitement simpliste éclaterait les cœurs similaires – mêmes rapports de taille – en une ribambelle de points... Si l'on s'intéresse véritablement à la *forme* du cœur, aux atrophies/hypertrophies éventuelles de ses ventricules, il faudra nécessairement travailler dans un espace quotient – sur lequel la notion de régression n'est a priori pas définie.

Création d'atlas et analyse statistique Autre axe de recherche, dans la même veine : l'étude statistique d'une population de formes. On le sait, l'étude approfondie d'un jeu d'indicateurs scalaires – poids, âge, tension artérielle... – peut nous apprendre beaucoup sur l'état de santé d'un patient. En comparant celui-ci à un humain moyen, en le replaçant au sein d'une population connue, on peut le classer dans tel ou tel groupe, détecter l'apparition de certaines pathologies. Alors, au XXI^{e} siècle, sera-t-il possible de faire la même chose avec des données plus complexes ? On pense par exemple à des images de colonne vertébrale (scoliose...), de fonds d'œils rétiniens (glaucome...) ou à la forme de l'hippocampe, corrélée à la présence de démences dégénératives comme la maladie d'Alzheimer.

Matching et transfert de données Enfin, un dernier type de problème issu de la neuro-imagerie : la création de cartes sémantiques sur les surfaces corticales de sujets observés. Pensons par exemple à une collection de visages. Il est aisé pour un être humain de "découper", *segmenter* ceux-ci en régions anatomiques bien définies : nez, yeux, bouches, oreilles... Toutes sont d'apparences, de topologies bien différenciées : il est donc possible d'apprendre à un ordinateur à reconnaître ces différents point saillants sur une image.

Mais comment procéder lorsque "toutes les régions se ressemblent" ? Sur une simple image anatomique, difficile en effet de distinguer les circonvolutions du cortex visuel de celles du lobe frontal... On s'efforcera donc de transférer une *carte* – dessinée par des experts – d'un cerveau *modèle* typique à un cerveau donné, de manière anatomiquement crédible – voir Figure 6.13.

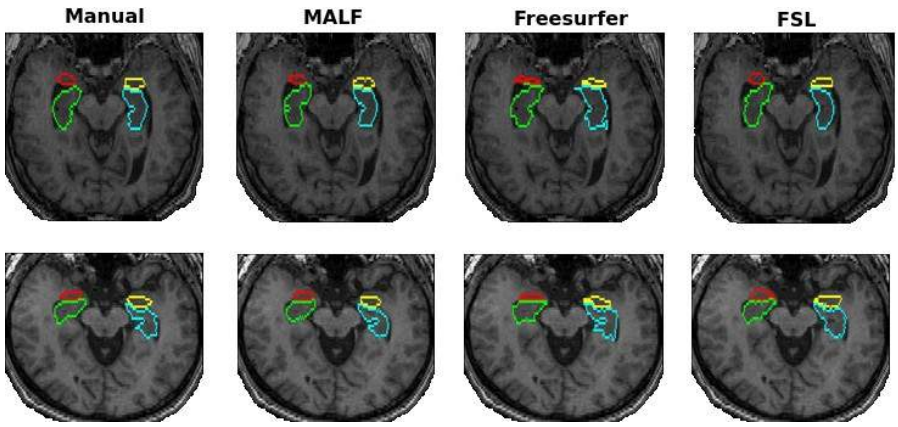
Interprétations métriques des outils statistiques usuels Dans les trois situation précédentes, tout serait simple si l'on disposait d'une structure algébrique/vectorielle raisonnable sur nos données – qui sont ici des "formes" – : moyenne, régression linéaire et analyse de la variance de l'échantillon sont toujours données par des formule pleines de « + » et de « × »... Mais malheureusement, une théorie additive des formes a peu de chance de voir le jour : quel sens pourrait-on donner à la somme de deux cœurs ?

Qu'à cela ne tienne : une notion de *distance* comme définie dans la section précédente suffira. La *moyenne* ne sera plus qu'un point minimisant la somme des distances aux observations et la droite de *régression*, une courbe géodésique approchant au mieux les données aux instants d'observation.

Résultats pratiques Cette dernière remarque légitime tout notre travail. Dans les pages qui suivent, on donne quelques résultats probants obtenus à l'aide de techniques de pointe qui font d'ores et déjà partie du quotidien de nombreux chercheurs et neurologues. On démontre par là l'utilité d'un modèle de déformations plus riche que celui des seules similitudes.

Le domaine est encore jeune, et les perspectives de recherche sont nombreuses. Pour le mathématicien se posent en fait trois grandes questions : Comment construire des métriques locales *anatomiquement pertinentes* sur un espace de formes ? Quelles seront alors les *propriétés* de l'espace métrique induit – courbure, etc. ? Et surtout, sera-t-on capable d'*implémenter* de manière efficace les algorithmes de régression, matching, création d'atlas que nous demandent les médecins/neurologues/biologistes ?

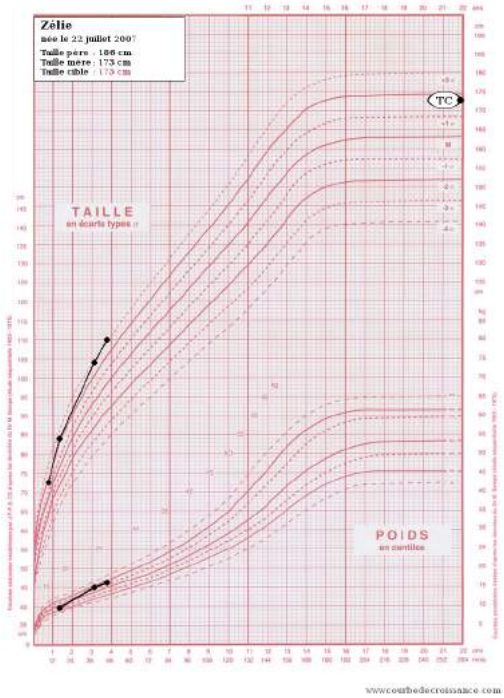
Nous n'irons pas plus loin dans ce cours de vulgarisation, qui touche à sa fin ; au lecteur intéressé, je suggère le polycopié écrit avec mes élèves (en première année au DMA) : *Introduction à la Géométrie Riemannienne par l'Étude des Espaces de Formes*, accessible à l'adresse : `www.math.ens.fr/~feydy/Teaching/geometrie_riemannienne_espaces_de_formes.pdf`.



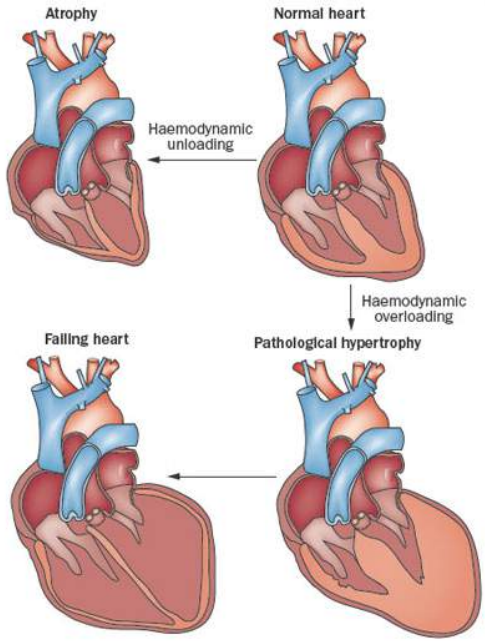
(a) Exemples de segmentation de régions du cerveau. Deux sujets sont observées, quatre méthodes sont utilisées pour transporter une carte pré-établie par la communauté scientifique.
Image tirée de *Segmentation of brain magnetic resonance images based on multi-atlas likelihood fusion : testing using data with a broad range of anatomical and photometric profiles*, Tang et Al., Frontiers in Neurosciences, 03 March 2015.

(b) À partir d'un modèle de référence, on peut inférer les mécanismes internes de silhouettes arbitraires : Le David, Olive, Brutus, un loup-garou... Des débouchés dans les domaine de l'imagerie médicale et de l'animation sont envisagés : vous pourrez trouver des vidéos de présentation sur YouTube aux adresses `www.youtube.com/watch?v=HgvDfQB4ajA` et `www.youtube.com/watch?v=ddp996DIZOk`.
Image tirée de *Anatomy Transfer*, Dicko et Al., ACM Transactions on Graphics, 2013.

FIGURE 6.13 – Deux exemples de transfert de modèle anatomique, qui tirent parti des *déformations de l'espace ambiant* produites par la plupart des méthodes de recalage d'images. En mettant en correspondance deux scans IRM ou deux silhouettes, on peut transporter sur l'image d'arrivée une information *a priori* connue dans l'espace de départ, comme une carte de segmentation ou un atlas anatomique.

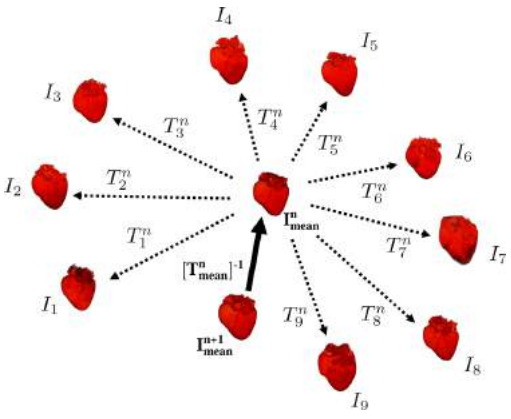


(a) Courbe de croissance, qui permet de comparer un enfant à l'ensemble de la population.

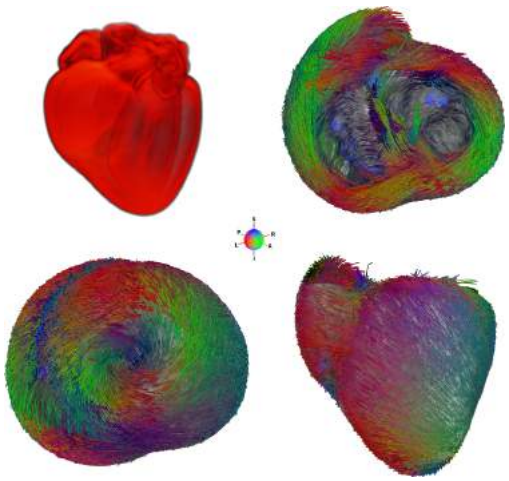


(b) Exemples de malformations cardiaques. Comment les détecter avant qu'il ne soit trop tard ?

FIGURE 6.14 – Le suivi longitudinal et les techniques de régression méritent d'être généralisés aux formes. (b) est tiré de *Vascular endothelial growth factor in heart failure*, Taimeh et Al., Nature Reviews Cardiology 10 (Septembre 2013).

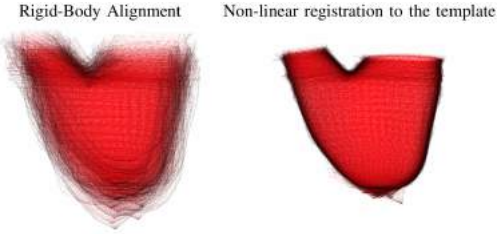


(a) Construction de l'atlas.

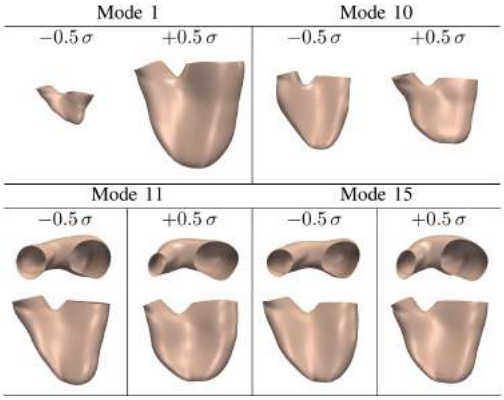


(b) Cœur moyen, avec son mille-feuilles de fibres musculaires anisotropes.

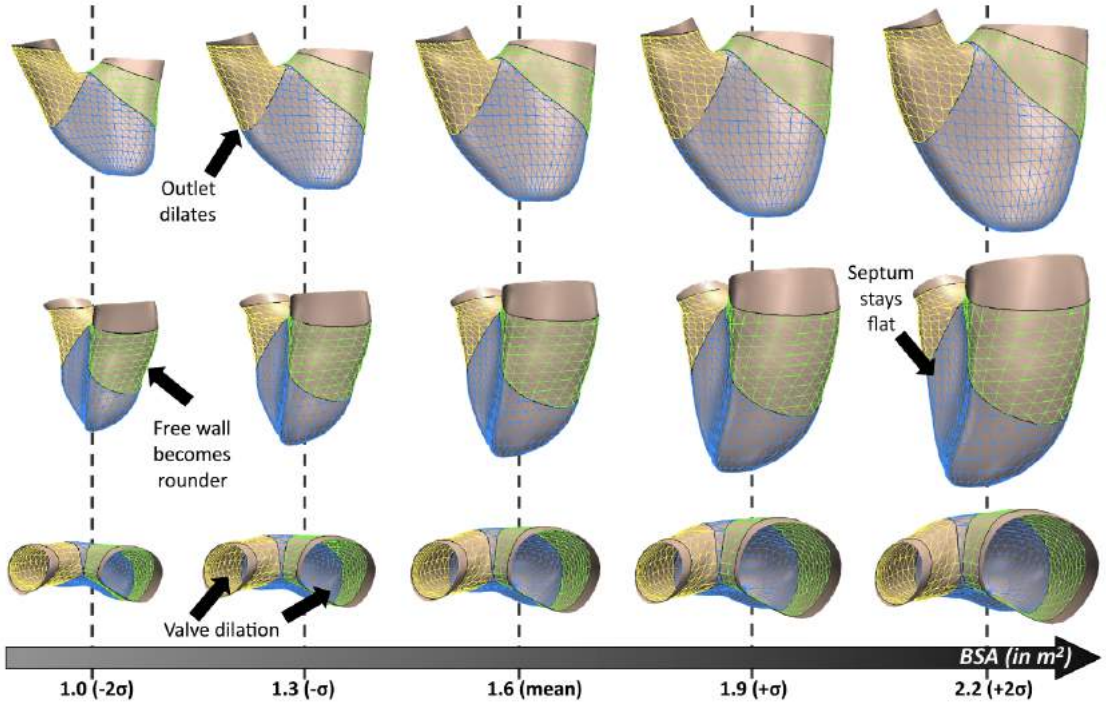
FIGURE 6.15 – Estimation d'un cœur moyen à partir de sept cœurs de chiens.
Images tirées de *A computational framework for the statistical analysis of cardiac diffusion tensors : application to a small database of canine hearts*, Peyrat et Al., IEEE transactions on medical imaging, 2007.



(a) Recalage des données sur la moyenne.

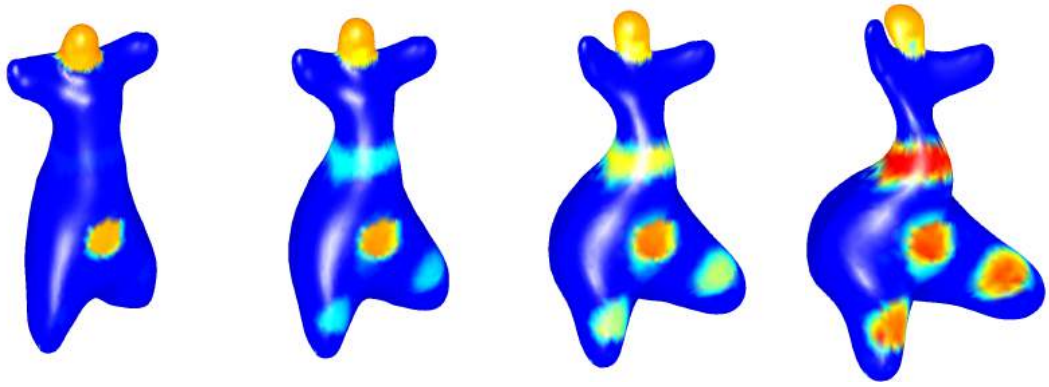


(b) Quelques modes de déformation typiques.



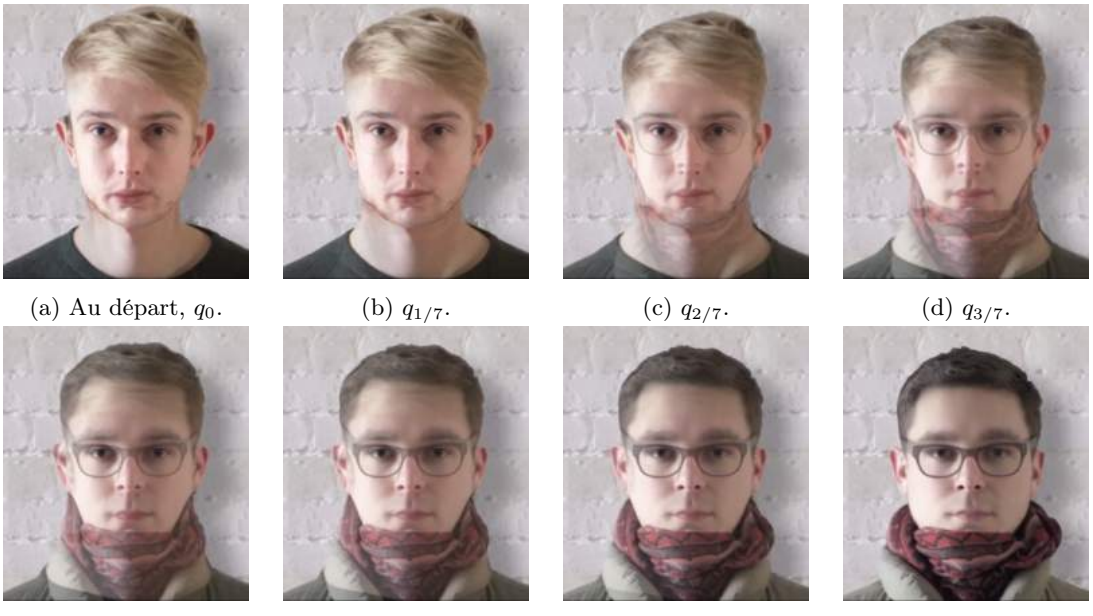
(c) Estimation des déformations typiques du template selon la "Body Surface Area".

FIGURE 6.16 – Autre estimation d'atlas de données cardiaques. Ici, les jeu de données est constitué de surfaces de ventricules droits (a, gauche). Par une analyse de formes analogue (mais pas identique) à celle présentée dans ce chapitre, les auteurs ont pu estimer une moyenne (a, droite) et des champs de vecteurs encodant les déformations de celle-ci vers les observations. Une analyse statistique classique peut alors être conduite dans l'espace tangent à la moyenne, qui est un espace vectoriel (muni d'une structure additive, etc.) encodant les déformations du *template*. Les principaux modes de déformation sont extraits (b) : on trouve par exemple que le "Mode 1" (principal) correspond à une dilatation/rétraction du cœur. In fine, on peut tracer des courbes d'analyse fine, qui donnent la corrélation typique entre surface corporelle et forme du cœur (c), avec en vue la détection d'une hypertrophie cardiaque, la tetralogie de Fallot.
Images tirées de *A statistical model for quantification and prediction of cardiac remodelling : Application to tetralogy of fallot*, Mansi et Al., IEEE transactions on medical imaging, 2011.



(a) Au départ, q_0. (b) $q_{1/3}$. (c) Au départ, $q_{2/3}$. (d) À l'arrivée, q_1.

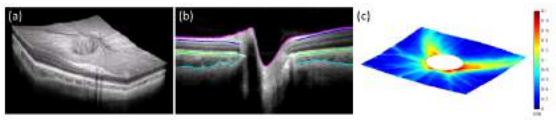
FIGURE 6.17 – Exemple de *métamorphose*, ou *déformation géometrico-fonctionnelle*, le long d'une trajectoire géodésique : cette double-page est consacrée aux applications pratiques d'outils d'analyse de fonctions (bidimensionnelles) combinant variation du signal et déformation du support. Images tirées de la thèse de Nicolas Charon, *Analysis of geometric and functional shapes with extensions of currents : applications to registration and atlas estimation.*



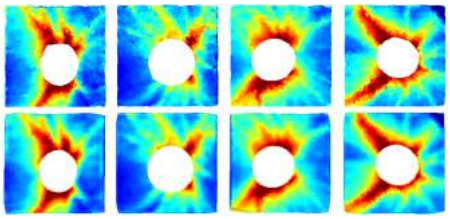
(a) Au départ, q_0. (b) $q_{1/7}$. (c) $q_{2/7}$. (d) $q_{3/7}$.

(e) Au départ, $q_{4/7}$. (f) $q_{5/7}$. (g) $q_{6/7}$. (h) À l'arrivée, q_1.

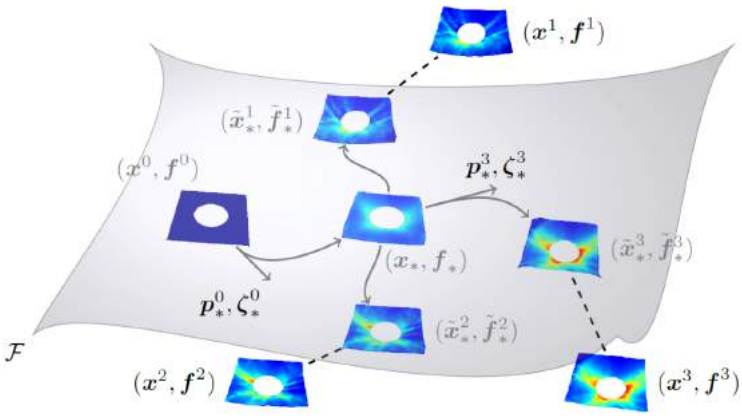
FIGURE 6.18 – Métamorphose entre deux visages, tirée du clip *Matilda* du groupe Alt-J (Δ) – `www.youtube.com/watch?v=Q06wFUi5OM8`. Observez comme des features géométriques (forme des oreilles, coiffure) sont continûment déformées, tandis que les variations fonctionnelles (lunettes, foulard) sont réalisées par modification de l'image, via un simple fondu. Un matching "par métamorphose" permet bien de distinguer ces deux informations et, in fine, de comparer des images définies sur des supports géométriques différents : variation *anatomique* assimilée à la composante *géométrique* de la métamorphose ; variation de *style vestimentaire* identifiée à sa partie *fonctionnelle*.
Notez que ce morphing n'est pas totalement satisfaisant (je ne sais pas comment il a été obtenu, probablement par un étiquetage automatique ou manuel des points saillants du visage + déformations localement affines) : le changement de nez est ici compris comme une information fonctionnelle, alors que nous voudrions l'inclure dans la partie géométrique.



(a) Pré-traitement des données.



(b) Sur la première ligne, quatre données typiques (q^i, f^i). Sur la seconde, les *modèles* $(\tilde{x}^i_*, \tilde{f}^i_*)$ correspondants.



(c) Construction d'un atlas conforme au modèle de déformation $\mathcal{F}$ à partir de données géometrico-fonctionnelles (x^i, f^i).

FIGURE 6.19 – Un jeu de données "cas d'école", collection de fonds d'œils rétiniens obtenus par tomographie à cohérence optique. L'objectif est d'arriver à une détection automatique du glaucome, pathologie causée par une surpression intra-oculaire qui entraîne asphyxie et dégénérescence progressive de la rétine autour du point de branchement du nerf optique (le "trou" dans nos images), selon des patterns qui correspondent à la vascularisation sous-jacente (variable selon les individus).
Les données brutes sont traitées en (a) : plutôt que de travailler avec un mille-feuille de couches rapprochées, on préférera assimiler les rétines à des feuillets simples dans l'espace, sur lesquels un signal (une *fonction*) donne l'épaisseur en microns.
Dans notre analyse de population, on veut séparer la variation *géométrique*, qui détermine les positions des vaisseaux sanguins autour du nerf optique et par là le support des segments rouges de (b), d'une variation *fonctionnelle*, la force ou l'atténuation du signal au dessus de ces courbes. À terme, on pourrait classifier les images en un groupe "glaucome" et un groupe "test" en regardant cette seule information fonctionnelle indépendante de la vascularisation sous-jacente.
Pour cela, on construit un atlas figuré en (c), donnée d'une *moyenne* (x_*, f_*) et de *modèles* déformés $(\tilde{x}^i_*, \tilde{f}^i_*)$ qui approchent au mieux les données réelles (x^i, f^i). Ici, la template est obtenue par déformation d'une forme de départ (x^0, f^0) et est choisie de manière à minimiser la somme des écarts aux données.
La variété $\mathcal{F}$ des déformations de la template est analogue à la variété $\mathcal{M}$ des déformations de nuages de points : on peut caractériser les modèles $(\tilde{x}^i_*, \tilde{f}^i_*)$ par des *moments de tir* (p^i_*, ξ^i_*) portés par la template, sur lesquels une analyse vectorielle classique (analyse en composantes principales...) est possible.
On pourra trouver une vidéo illustrant la création de l'atlas sur la page personnelle de Benjamin Charlier, `www.math.univ-montp2.fr/~charlier/soft/videos/atlas_H1.webm`.
Ces images sont tirées d'un article reposant explicitement sur l'algorithme LDDMM présenté dans ce chapitre, *Atlas-based Shape Analysis and Classification of Retinal Optical Coherence Tomography Images using the Functional Shape (fshape) Framework*, par Sieun Lee, Nicolas Charon, Benjamin Charlier, Karteek Popuri, Evgeniy Lebed, Marinko V. Sarunic, Alain Trouvé et Mirza Faisal Beg.

Le travail du mathématicien appliqué

On l'a vu, permettre aux biologistes, neurologues et autres radiologues d'aller au delà de la simple analyse procustéenne nous aura demandé de faire appel à des concepts de géométrie Riemannienne tout à fait non-triviaux. Pour faire du bon travail, un effort de *communication* est donc indispensable, sous tous les plans.

D'abord, parce qu'il est essentiel de se tenir au courant des nouvelles idées, des méthodes concurrentes qui font avancer l'état de l'art et la compréhension du problème. Mais aussi, tout simplement, parce qu'être à l'écoute de nos "utilisateurs finaux" nous permet d'envisager de nouvelles questions qui, en plus d'être pertinentes, peuvent nous mener à ouvrir de nouvelles portes. Plutôt que d'écrire sur du vent, pourquoi ne pas vous présenter mes propres sujets de recherche en anatomie computationnelle ?

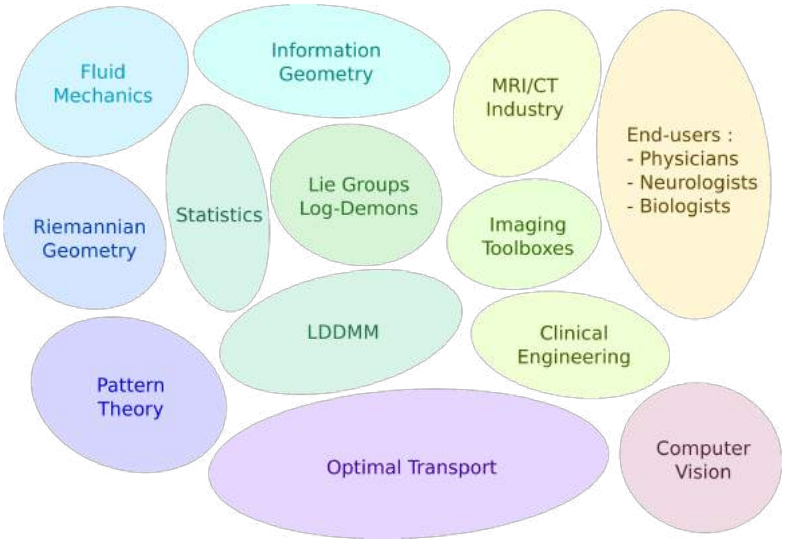
Difféomorphométrie de données partiellement observées D'abord, un premier problème : celui de l'adéquation aux problèmes médicaux d'un modèle de déformations seules. En effet, celles-ci ne permettent pas de prendre en compte les problèmes de *bords* induits par les *cadrages* des images au moment de l'acquisition, ou l'existence de données lacunaires. Par exemple, dans le cas de scans IRM post-AVC, les régions du cerveau mortes disparaissent et sont absentes des images : comment, alors, pourrait-on recaler un cerveau moyen sur ces individus pleins de "trous" ?

Transport optimal corrélé Théorie LDDMM et transport optimal sont liés... Mais jusqu'où peut-on pousser l'analogie ? Pourrait-on trouver un cadre théorique, algorithmique et pratique qui permette de lier la continuité de la théorie LDDMM avec la sensibilité aux masses du transport optimal ?

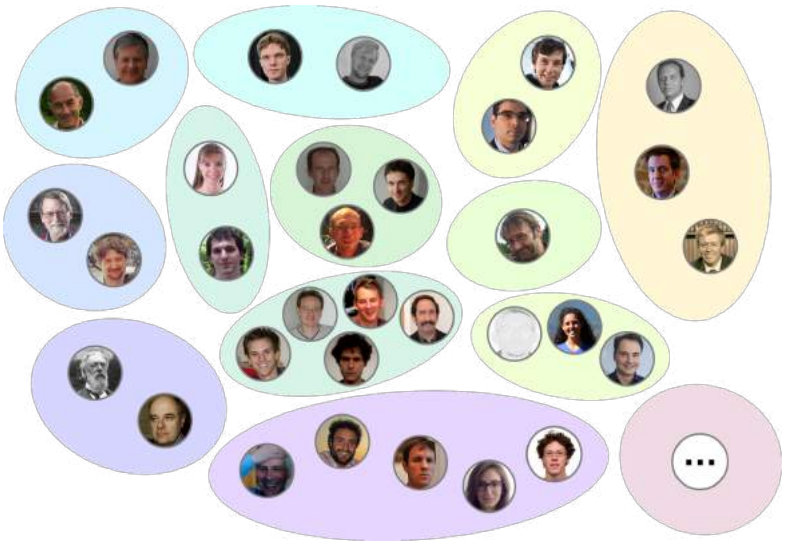
Éventail des domaines, des profils Mes recherches tournent autour de ces thèmes, auxquels on peut ajouter une problématique plus technique liée à l'estimation de sous-espaces de régression optimaux. Tous sont liés à des discussions que j'ai pu avoir avec des médecins, des ingénieurs, ou tout simplement des collègues mathématiciens. C'est que dans son travail, un mathématicien appliqué est toujours amené à interagir avec des spécialistes de domaines divers, qui vont dans notre cas de la géométrie fondamentale à la radiologie clinique – voir Figure 6.20.

Conclusion Dans ce cours, j'ai toujours voulu mettre en avant le cœur du travail mathématique : la recherche de *représentations adaptées*. J'espère vous avoir aidé à porter un regard nouveau sur la logique fondamentale, les nombres complexes ou la dimension infinie... Mais, surtout, je voudrais avoir démystifié à vos yeux notre profession ; si éloignée du grand public, et pourtant tellement influente, par petites touches, sur son quotidien.
Qu'en avez-vous pensé ? Un livre, donc ce polycopié fournira la trame, est en cours de préparation. Je serais absolument ravi d'entendre vos remarques et critiques avisées : n'hésitez donc pas à m'écrire à mon adresse mail professionnelle, `jean.feydy@ens.fr`.

D'ici là, portez-vous bien !



(a) Les principaux domaines liés à la théorie LDDMM.



(b) Derrières les étiquettes, des chercheurs.

FIGURE 6.20 – Avant d'être écrites dans les manuels, les mathématiques sont développées, pensées par une large communauté de spécialistes. De haut en bas et de gauche à droite (du plus fondamental au plus appliqué), on retrouvera : *Mécanique des fluides* : Vladimir Arnold, Peter Michor ; *Géométrie Riemannienne* : David Mumford, Mario Micheli ; *Théorie des Patterns* : D'Arcy Thompson, Ulf Grenander ; *Géométrie de l'information* : Martins Bruveris, Martin Bauer ; *Statistiques* : Stéphanie Alassonnière, Loïc Devilliers ; *Log-demons* : Xavier Pennec, Nicolas Ayache, Marco Lorenzi ; *LDDMM* : Nicolas Charon, Laurent Younès, Joan Glaunès, Alain Trouvé, Michael Miller ; *Transport optimal* : Yann Brenier, Filippo Santambrogio, Gabriel Peyré, Aude Genevay, Lénaïc Chizat ; *Industrie de l'imagerie médicale* : Tommaso Mansi, Julian Krebs ; *Toolbox d'imagerie* : John Ashburner ; *Liens Maths/Clinique* : Benjamin Charlier, Barbara Gris, Stanley Durrleman ; *Utilisateurs finaux* : Alexandre Krainik, Antoine Feydy, Karl Friston.
Bien sûr, cette liste est loin, très loin d'être exhaustive... Mais au travers de ce petit échantillon, vous pouvez déjà apprécier combien l'avancée des mathématiques et sa transition vers les applications concrètes nécessite de travail. Si vous êtes curieux, pourquoi ne pas visiter les pages web personnelles de tout ces chercheurs ? Vous pourrez ainsi vous faire une bonne idée du continuum de profils impliqués dans cette marche vers le progrès.

Deuxième partie

Fondements des mathématiques : de la logique aux distributions

Chapitre 7

Preuves formelles, axiomatiques et théorie des ensembles

Séances 8 et 9

Revoir l'intro.

Quelle est la nature d'un nombre, d'un objet mathématique ? Cette question, si âprement disputée à la fin du XIX^{e} siècle, n'intéresse plus guère la communauté mathématique. Et pour cause : le travail de longue haleine mené par des logiciens tels que Frege, Russel ou Gödel a permis d'arriver aux surprenantes conclusions que voici – elles seront précisées et développées tout au long du chapitre :

1. Pour faire des mathématiques, la *nature* des objets étudiés n'a pas d'importance. Seules comptent les *relations* entre ceux-ci qui peuvent, dans tous les cas intéressants, se réduire à un petit nombre de règles du jeu élémentaires, les *axiomes* d'une théorie.

2. Si une théorie est *cohérente*, c'est à dire s'il n'est pas possible d'arriver à une *contradiction* à partir de ses axiomes, alors il est possible d'en construire un modèle formel – et réciproquement. Autrement dit, pour une théorie, la cohérence est équivalente à l'existence d'une structure formelle vérifiant ses axiomes – dans la pratique, celle-ci est construite à partir de chaînes de caractères. C'est (presque) le théorème de complétude de Gödel.

3. Arithmétique, analyse, géométrie... Tous les domaines des mathématiques peuvent s'écrire, se modéliser dans le langage de la *théorie des ensembles* muni du jeu d'axiomes "ZF", via des constructions formelles que l'on pourrait assimiler à des "émulateurs" ; on parle d'*encodage*. Si le jeu d'axiomes ZF est cohérent, alors tout le reste des mathématiques l'est aussi : on n'arrivera pas à prouver que "$0 = 1$", ou que "π est à la fois rationnel et irrationnel".

4. Clé de voûte de cette belle certitude, la *cohérence de la théorie ZF*, est indémontrable. Pour être plus exact : si l'on en trouve une démonstration, alors on peut en produire une autre qui démontre le contraire ; la théorie serait donc incohérente, puisqu'elle permet de démontrer une chose et son contraire.

Ce dernier résultat, le fameux *second théorème d'incomplétude de Gödel*, a mis un point final aux espoirs d'une théorie mathématique "auto-démontrée". S'il est possible de ramener la cohérence des théories mathématiques les plus extravagantes à celle d'un jeu d'axiomes extrêmement raisonnable (le fameux système ZF), le mathématicien est néanmoins contraint de faire acte de foi : *croire* en la cohérence de celui-ci, jusqu'à preuve du contraire.

Formules logiques

On l'aura compris, cet échafaudage logique permet de mettre de côté la question de la *nature* profonde des objets considérés : seules comptent les relations entre objets, formalisées en un jeu d'axiomes que l'on espérera être cohérent. Nous aborderons la question de la pertinence de cette réduction des mathématiques à un simple jeu formel au fil des prochaines séances... Mais, avant de nous aventurer sur ce terrain polémique, au-delà des mathématiques, précisons maintenant le portrait brossé en introduction du chapitre.

Pour "construire les mathématiques", on se repose sur la seule base solide acceptée à coup sûr par le lecteur : les caractères présents sur sa feuille. Ceux-ci, choisis dans un alphabet, ou *langage* donné, seront ensuite assemblés en *formules*.

Définition 7.1 (Langage – du premier ordre)**.** Un *langage* est une collection de symboles $\mathcal{L}$ qui se compose de deux parties :

1. La première partie (commune à tous les langages) consiste en les symboles auxiliaires « (» et «) » ainsi qu'en les *symboles logiques* suivants :
   - la collection de variables $\mathcal{V} = \{v_0, v_1, v_2, \dots\}$ (numérotées par commodité),
   - le symbole de l'égalité $=$ (« égal »),
   - les connecteurs $\neg$ (négation, « non »), $\wedge$ (conjonction, « et »),
   - le quanteur existentiel $\exists$ (« il existe »).

   On note cette partie fixe de notre langage $\mathcal{L}_{log}$.
2. La deuxième partie, variable, consiste en les *symboles non logiques* de $\mathcal{L}$. On y trouve :
   - une collection de constantes $\mathcal{C}^{\mathcal{L}}$,
   - une collection de symboles fonctionnels $\mathcal{F}^{\mathcal{L}}$,
   - une collection de symboles de relations $\mathcal{R}^{\mathcal{L}}$.

Un premier exemple Le langage des corps ordonnés (utilisé pour les opérations usuelles sur les nombres) :

$$\mathcal{L}_{ann} = \mathcal{L}_{log} \cup \{\underbrace{0,1}_{\mathcal{C}}, \underbrace{+,\times}_{\mathcal{F}}, \underbrace{<}_{\mathcal{R}}\}. \tag{7.1}$$

$+$, $\times$ sont ici des symboles *fonctionnels* binaires, ce qui signifie que $+(a,b)$, que l'on abrège en $(a+b)$, s'utilise syntaxiquement comme un *élément*. À l'inverse, $<$ est ici un symbole de *relation* binaire, ce qui signifie que $<(a,b)$ – que l'on allège en $(a<b)$ – s'utilise syntaxiquement comme un *booléen*, une valeur logique. Enfin, 0 et 1 sont des constantes remarquables, utilisables comme des éléments. Les inclure explicitement dans notre langage est une commodité, pas véritablement nécessaire.

Règles de composition Il est bien entendu possible de formaliser les règles de composition d'une formule *bien écrite*... Mais je pense que vous préférerez une petite liste d'exemples, assortie d'une pincée de bon sens ! Saurez-vous lire les formules ci-dessous, rédigées sur le langage $\mathcal{L}_{ann}$?

$(1<0)$	bien écrite	(7.2)
$1<<(0+0)$	mal écrite	(7.3)
$\exists\, x,\, x<0$	bien écrite, *fermée*	(7.4)
$\exists\, x,\, y<0$	bien écrite, mais *ouverte*	(7.5)
$\neg\Big[\,\exists\, x, \neg\Big(\,\neg\big(\,\neg\,(\,\neg\,(0<x)\,)\wedge\neg\,(\,x<((1+1)\times x)\,)\,\big)\,\Big)\,\Big]$	bien écrite, *fermée*	(7.6)

Raccourcis d'écriture Cette dernière formule est bien fastidieuse. Pour conserver une certaine intelligibilité, on ajoute à notre langage un "sucre syntaxique" défini comme suit :

« pour tout » :	$\forall\, x,\ \varphi$	sera un raccourci pour	$\neg\,(\exists\, x,\ \neg\,\varphi)$	(7.7)
« ou » :	$\varphi \vee \psi$	sera un raccourci pour	$\neg\,\big((\neg\,\varphi) \wedge (\neg\,\psi)\big)$	(7.8)
« implication » :	$\varphi \Rightarrow \psi$	sera un raccourci pour	$(\neg\,\varphi) \vee \psi$	(7.9)
« équivalence » :	$\varphi \Leftrightarrow \psi$	sera un raccourci pour	$(\varphi \Rightarrow \psi) \wedge (\psi \ \Rightarrow \varphi)$.	(7.10)
				(7.11)

La formule (7.6) sera donc syntaxiquement équivalente à l'énoncé plus simple :

$$\forall\, x,\ \neg\Big(\neg\,(\neg\,(0 < x)\,) \wedge \neg\,(\,x < ((1+1) \times x)\,)\Big), \tag{7.12}$$

puis à

$$\forall\, x,\ \neg\,(0 < x) \vee (\,x < ((1+1) \times x)\,), \tag{7.13}$$

$$\forall\, x,\ (0 < x) \Longrightarrow (\,x < ((1+1) \times x)\,). \tag{7.14}$$

Axiomatiques et vérités

Démonstrations formelles

On sait maintenant rédiger de jolies formules mathématiques... Mais comment leur assigner une valeur de validité, de vérité ? Il faut commencer par choisir une collection arbitraire de formules sur le langage $\mathcal{L}$,

$$\mathrm{Ax} = \{A_1, A_2, \ldots, A_p\}, \tag{7.15}$$

étiquetées comme *Vraies*. On dira que Ax est un *jeu d'axiomes* sur $\mathcal{L}$. Il existe alors deux manières a priori concurrentes d'assigner une valeur de vérité aux formules sur $\mathcal{L}$.

Preuves finies La première – que vous connaissez bien – consiste à enchaîner des énoncés les uns à la suite des autres par des règles de logique élémentaire, échafaudant ainsi une *démonstration* :

Définition 7.2 (Axiomes logiques)**.** Comme pour les langages avec la sous-collection $\mathcal{L}_{log}$, il y a une série d'axiomes logiques notée Ax_{log}, que l'on souhaite ajouter à toute théorie, et que l'on omettra donc de préciser par la suite :

1. Les tautologies, c'est à dire les formules obtenues en substituant des sous-formules $\psi_1, \ldots, \psi_n$ dans une formule booléenne $F = F(q_1, \ldots, q_n)$ valant 1 pour toute affectation des q_i à 0 ou 1.
   Par exemple, à partir de la formule booléenne

   $$F(q_1) = q_1 \vee (\neg\, q_1) \tag{7.16}$$

   et de la formule $\psi_1 : \forall\, x, 0 < x$, on obtient la tautologie

   $$(\forall\, x, 0 < x) \vee \neg\,(\forall\, x, 0 < x). \tag{7.17}$$

   C'est le principe du tiers exclu.

Revoir l'intro.

2. Les axiomes de l'égalité :

$$\forall x,\ x = x \qquad \text{(réflexivité)} \tag{7.18}$$
$$\forall x,\ \forall y,\ x = y \Rightarrow y = x \qquad \text{(symétrie)} \tag{7.19}$$
$$\forall x,\ \forall y,\ \forall z,\ (x = y \wedge y = z) \Rightarrow (x = z) \qquad \text{(transitivité)} \tag{7.20}$$

qui garantissent que « = » se comporte comme une *relation d'équivalence* ; il faut y ajouter les axiomes suivants – un pour chaque symbole fonctionnel n-aire f de $\mathcal{F}^{\mathcal{L}}$ et chaque symbole relationnel m-aire R de $\mathcal{R}^{\mathcal{L}}$ – qui garantissent qu'application de fonction et comparaison par une relation passent bien à l'égalité :

$$\forall x_1, \dots, x_n,\ \forall y_1, \dots, y_n,\ (x_1 = y_1 \wedge \dots \wedge x_n = y_n) \tag{7.21}$$
$$\Longrightarrow f(x_1, \dots, x_n) = f(y_1, \dots, y_n) \qquad \text{(congruence fonctionnelle)}$$
$$\forall x_1, \dots, x_m,\ \forall y_1, \dots, y_m,\ (x_1 = y_1 \wedge \dots \wedge x_m = y_m \wedge R(x_1, \dots, x_m)\,) \tag{7.22}$$
$$\Longrightarrow R(y_1, \dots, y_m) \qquad \text{(congruence relationnelle)}$$

3. Les axiomes du quanteur existentiel : pour toute formule φ, pour tout *terme* t – un mot obtenu par application de fonctions sur les constantes et les variables, i.e. une formule sans relations ni quanteur –, on ajoute l'axiome

$$\varphi_{\text{où } t \text{ remplace } x} \Longrightarrow \exists\ x,\ \varphi. \tag{7.23}$$

Par exemple, avec $\varphi : (\forall\ y,\ y < x)$ et $t : (1 + 1) \times (0 + 1)$, on obtient l'axiome

$$\forall\ y,\ y < (1 + 1) \times (0 + 1) \Longrightarrow \exists\ x,\ \forall\ y,\ y < x. \tag{7.24}$$

Définition 7.3 (Règles de déduction)**.** Les maillons de nos preuves sont liés par :

1. Le Modus Ponens : À partir de φ et $\varphi \Rightarrow \psi$, on déduit ψ.
2. L'introduction du quanteur existentiel : Si x n'apparaît pas dans ψ, à partir de $\varphi \Rightarrow \psi$, on déduit $\exists\ x, \varphi \Rightarrow \psi$.

Définition 7.4 (Preuve formelle – finie)**.** Soit φ une $\mathcal{L}$-formule, Ax un jeu d'axiomes.
Une *preuve formelle* de φ dans Ax est une suite finie de $\mathcal{L}$-formules $(\varphi_0, \varphi_1, \dots, \varphi_n)$ telle que :

- $\varphi_n = \varphi$ – on termine la preuve par un CQFD ;
- pour tout $i \leqslant n$ on a :
  - ou bien φ_i est dans Ax ou Ax_{log} ;
  - ou bien φ_i s'obtient par Modus Ponens à partir de φ_j, φ_k avec $j, k < i$;
  - ou bien φ_i s'obtient par introduction du quanteur existentiel à partir d'une formule φ_j avec $j < i$.

On dira que φ est *prouvable* dans Ax s'il existe une preuve formelle de φ dans Ax.

Attention : les axiomes *logiques* et les règles de déduction données ci-dessus n'ont rien d'*intrinsèquement* « Vrai ». Les utiliser relève d'une *décision* prise par les mathématiciens : certains logiciens rejettent par exemple l'usage du tiers exclu, considéré comme non-légitime. La logique mathématique classique, qui étudie ce jeu symbolique particulier, n'a donc rien d'*absolu* au sens philosophique : sa légitimité viendra du théorème 7.1 (complétude de Gödel), qui affirme – dans un contexte méta-mathématique extrêmement raisonnable – que tout énoncé "conséquence nécessaire" d'un jeu d'axiomes est démontrable à partir de celui-ci et des axiomes *logiques* via l'utilisation des règles de déduction.

Exemple fondamental : La théorie des ensembles

Avant d'aller plus loin, il me semble judicieux d'étoffer nos définitions générales de quelques exemples concrets. Plutôt que de travailler avec le langage et les axiomes de la théorie des *groupes*, des *anneaux* et autres structures algébriques à base de $+$ et de $\times$, penchons-nous sur la théorie des *ensembles*.

Qu'est-ce qu'un ensemble ? Intuitivement, une collection d'objet, une idéalisation mathématique de la "boîte"... Ce qui est plutôt vague ! À vrai dire – comme le dit très bien Patrick Dehornoy –, faute de pouvoir définir commodément les ensembles à partir d'objets plus primitifs, on se contentera d'une approche *axiomatique* : les ensembles seront les variables du langage $\mathcal{L}_{ens}$ défini pour l'occasion,

$$\mathcal{L}_{ens} = \mathcal{L}_{log} \cup \{ \in \}, \tag{7.25}$$

où le seul symbole non-logique « $\in$ », « appartient », est un symbole de relation binaire.

Vient alors la question de savoir quels axiomes prendre pour constituer notre théorie. Le premier d'entre eux, indiscutable, sera le suivant :

$$\forall\, x,\ \forall\, y,\ \big(\ (\ \forall\, z, (\, z \in x \Leftrightarrow z \in y)\,) \Longrightarrow x = y\ \big) \qquad \text{(extensionalité)} \tag{7.26}$$

Autrement dit, deux ensembles qui ont les mêmes éléments sont égaux : on parle bien de collections abstraites, dépourvues de qualités propres comme la forme, la couleur, etc.

Qu'est-ce qui *est* un ensemble ? Mais comment choisir les axiomes suivants ? Via l'introduction de quanteurs d'existence « $\exists$ », ceux-ci détermineront en pratique *ce qui est* un ensemble.

La première tentative fut l'œuvre de Georg Cantor et son axiome de *compréhension* est le suivant : toute propriété P caractérise l'ensemble x_P des "éléments" qui la vérifient. Malheureusement, cette approche trop naïve s'écroule face au paradoxe de Berry :

Proposition 7.1 (Paradoxe de Berry)**.** *Soit $\mathcal{P}(n)$ la propriété :*
« n est un entier définissable par une phrase française d'au plus cent caractères.»
Alors il ne peut exister "d'ensemble $x_{\mathcal{P}}$ des entiers vérifiant $\mathcal{P}$".

Exercice 7.1. Trouver pourquoi !

Comment répondre à ce paradoxe ? On ne peut décemment renoncer à construire des ensembles d'entiers. Le choix des informaticiens serait de refuser toute définition "via une propriété", et d'accepter seulement les ensembles "construits" explicitement par unions, intersections de briques élémentaires... Ce qui est beaucoup trop restrictif pour un mathématicien ! Le compromis naturel, proposé par Gottlob Frege, est de restreindre le champs des propriétés "admissibles" à celles qui sont exprimables par une formule sur le langage $\mathcal{L}$ considéré ; on pense par exemple à

$$\mathcal{P}(n)\ : \exists\, k,\ n = (1+1) \times k \qquad \text{« être un nombre pair »}, \tag{7.27}$$

$$\mathcal{Q}(x)\ : \exists\, y,\ y \in x \qquad \text{« être non vide »}. \tag{7.28}$$

À vrai dire, le paradoxe de Berry ne faisait que révéler la confusion qui a longtemps régné autour de ces problématiques d'*existence*. On pense par exemple au célèbre argument ontologique : « Dieu a toutes les propriétés, donc il existe ». Aujourd'hui, cet énoncé qui identifie l'existence d'un objet – qui est une caractéristique du modèle : ici, l'univers – à une propriété portant sur les éléments de celui-ci passerait pour une erreur de logique élémentaire – sans graves conséquences, la foi religieuse étant plus intéressée par les mystères que par sa cohérence interne. Il a pourtant été

défendu par Descartes, un esprit dont la rigueur mathématique a tant frappé ses contemporains qu'elle en est devenue proverbiale !

L'approche de Frege, innovante et formelle, a donc le mérite de mettre les points sur les i en précisant les différents niveaux de discours... Elle ne résistera pourtant guère plus longtemps que celle de Cantor.

Proposition 7.2 (Paradoxe de Russel)**.** *Soit $\mathcal{E}$ la propriété*

$$\mathcal{E}(x) \ : \neg(x \in x) \qquad \text{« } x \text{ n'est pas élément de lui-même».} \tag{7.29}$$

Alors l'existence d'un ensemble $x_{\mathcal{E}}$ des ensembles vérifiant $\mathcal{E}$ est une hypothèse contradictoire.

Exercice 7.2. Comprendre pourquoi !

Il existe donc des propriétés sur les ensembles – les variables de notre langage – qui ne sauraient être *représentées* par l'un d'eux. L'approche classique pour répondre à ce (dernier) paradoxe est de l'attribuer au fait que l'*ensemble de tous les ensembles* est un objet trop grand pour être un ensemble... Une propriété $\mathcal{P}$ ne saura donc générer un ensemble qu'à travers la *restriction* d'un ensemble pré-existant.

Axiome de séparation, conséquences Pour toute formule $\varphi = \varphi(v_0, \dots, v_n)$ sur $\mathcal{L}_{ens}$, on ajoutera l'axiome de séparation associé :

$$\forall\, v_1, \dots, v_n, \forall\, a, \exists\, b, \forall\, x, \Big(x \in b \Leftrightarrow \big(x \in a \wedge \varphi(x, v_1, \dots, v_n)\big)\Big) \qquad \text{(axiome de séparation).} \tag{7.30}$$

Autrement dit : pour tout ensemble a, pour tous paramètres $v_1, \dots, v_n$, la formule

$$b = \{x \in a \text{ tels que } \varphi(x, v_1, \dots, v_n) \text{ est vérifié}\} \tag{7.31}$$

définit bien un ensemble.

Exercice 7.3. Montrer que le *schéma* d'axiomes de séparation légitime l'existence :

- de l'intersection $a \cap b$ de deux ensembles a et b ;
- de la différence $a \backslash b$ de deux ensembles a et b ;
- de l'ensemble vide $\emptyset$; quid de son unicité ?

Exercice 7.4. Montrer que le *schéma* d'axiomes de séparation permet de démontrer le théorème "de Russel" suivant :

$$\neg \exists\, x, \forall\, z, z \in x. \tag{7.32}$$

Création d'ensembles L'axiome de *séparation* proposé par Ernst Zermelo pour parer au paradoxe de Russel permet donc d'éviter de créer des ensembles trop gros... Et pour cause : agissant uniquement par restriction, il ne permet jamais d'augmenter le "nombre" d'éléments d'un ensemble ! Pour arriver à un cadre axiomatique suffisamment riche pour exprimer les objets mathématiques usuels, on convient donc de rajouter les axiomes suivants à notre théorie :

$$\forall\, a, b, \exists\, p, \forall\, x, \big(x \in p \Leftrightarrow (x = a \vee x = b)\big) \qquad \text{(axiome de la paire : } p = \{a, b\}\text{)}, \tag{7.33}$$

$$\forall\, a, \exists\, u, \forall\, x, \big(x \in u \Leftrightarrow \exists\, v, (x \in v \wedge v \in a)\big) \qquad \text{(axiome de la réunion : } u = \bigcup a\text{)}, \tag{7.34}$$

$$\forall\, a, \exists\, p, \forall\, x, \big(x \in p \Leftrightarrow (\forall\, t, t \in x \Rightarrow t \in a)\big) \qquad \text{(axiome des parties : } p = \mathcal{P}(a)\text{)}. \tag{7.35}$$

Exercice 7.5. Exprimer en des termes mathématiques usuels les trois axiomes ci-dessus.
L'axiome de la réunion légitime le fait de considérer l'union ensembliste d'une collection quelconque d'ensembles a, ce qui est un peu plus que ce à quoi on s'attend au premier abord... Partant de l'axiome de la paire et de celui de la réunion, légitimer l'existence ensembliste de la réunion $a \cup b$ de deux ensembles a et b – i.e. l'union telle qu'elle est présentée au lycée.

Exercice 7.6 (Couples). Les paires $\{a, b\}$ obtenues via l'axiome éponyme ne sont pas ordonnées... Démontrer que :

$$\forall\ a, b, c, d,\ \{\ a, \{a, b\}\ \} = \{\ c, \{c, d\}\ \} \Leftrightarrow (a = c \wedge b = d). \tag{7.36}$$

La construction $(a, b) = \{\ a, \{a, b\}\ \}$ permet donc de représenter les paires ordonnées (ou *couples*) de manière satisfaisante.

Axiomes de remplacement Unions, intersections, paires, couples... Toutes les opération ensemblistes usuelles sont maintenant légitimées, à une exception près : l'application de fonctions formelles définies par une formule. C'est chose faite avec le *schéma* d'*axiomes de remplacement*, qui stipule que pour toute formule $\varphi = \varphi(x, y, v_1, \ldots, v_n)$ sur $\mathcal{L}_{ens}$, on ajoutera l'*axiome de remplacement* associé :

$$\forall\ d, \forall\ v_1, \ldots, v_n,\ \Big(\forall\ x,\ \forall\ y_1, y_2,\ \big(\varphi(x, y_1, v_1, \ldots, v_n) \wedge \varphi(x, y_2, v_1, \ldots, v_n)\big) \Rightarrow y_1 = y_2\Big) \tag{7.37}$$

$$\Longrightarrow \exists\ z, \forall\ y, \Big(y \in z \Leftrightarrow \exists\ x, \big(x \in d \wedge \varphi(x, y, v_1, \ldots, v_n)\big)\ \Big). \tag{7.38}$$

Sacrée formule ! Vous arriverez sûrement à la comprendre si vous considérez que :

— φ est une "fonction" formelle ;
— $v_1, \ldots, v_n$ sont ses n paramètres ;
— x est un antécédent ;
— y, son "image" ;
— d, son domaine de définition.

Moralement, $\varphi(x, y, v_1, \ldots, v_n)$ est donc une formule logique qui vaut *Vrai* si $y = \varphi_{(v_1,\ldots,v_n)}(x)$.

La première ligne traduit alors que l'on n'acceptera de travailler qu'avec des formules *fonctionnelles*, qui à un antécédent x associent au plus une image y – à paramètres $v_1, \ldots, v_n$ fixés.

L'*ensemble* z ainsi défini, image de d par $\varphi_{(v_1,\ldots,v_n)}$, peut alors s'écrire

$$z = \Big\{\ \varphi_{(v_1,\ldots,v_n)}(x),\ \text{pour } x \text{ décrivant } d\ \Big\}. \tag{7.39}$$

Exercice 7.7. Les formules suivantes sont-elles *fonctionnelles* ?

$$\alpha(x, y) :\ \forall\ z, z \in y \Leftrightarrow z = x \tag{7.40}$$

$$\beta(x, y, v) :\ \big(\forall\ z, z \in y \Leftrightarrow z \in x\big) \vee (y = v) \tag{7.41}$$

$$\gamma(x, y, v) :\ \forall\ z, z \in y \Leftrightarrow (z \in x \vee z = v) \tag{7.42}$$

$$s(x, y) :\ \forall\ z, z \in y \Leftrightarrow (z = x \vee z \in x) \tag{7.43}$$

Si c'est le cas, écrire – en termes ensembliste usuels – l'image par celles-ci de l'ensemble

$$d = \Big\{\{\ \}, \{1\}, \{0, 1, 2\}\Big\} \tag{7.44}$$

pour $v = \{2\}$.

Exercice 7.8 (Produit cartésien, Ensembles d'applications).
1. On a vu dans l'exercice 7.6 comment construire la paire (x, y) à partir de deux ensembles x et y. À l'aide du schéma d'axiomes de remplacements, justifier l'existence ensembliste du *produit cartésien*

$$X \times Y = \big\{ (x, y), \text{ pour } x \text{ décrivant } X \text{ et } y \text{ décrivant } Y \big\} \tag{7.45}$$

de deux ensembles X et Y.
2. Une *application* $f : X \to Y$ entre un ensemble de départ X et un ensemble d'arrivée Y est la donnée d'un *graphe fonctionnel* g, sous-ensemble de $X \times Y$ tel que

$$\forall\, x \in X, \underbrace{\big(\exists\, y \in Y, (x, y) \in g\big)}_{x \text{ a une image}} \wedge \underbrace{\big(\forall\, y_1, y_2 \in Y, \big((x, y_1) \in g \wedge (x, y_2) \in g \big) \Rightarrow y_1 = y_2 \big)}_{\text{celle-ci est uniquement définie}}. \tag{7.46}$$

À l'aide de l'axiome des parties, de la question 1 et du schéma d'axiomes de séparation, montrer que l'ensemble Y^X des graphes fonctionnels de X vers Y est bien défini.

Axiomatique de Zermelo-Fraenkel Restent encore deux autres axiomes, sans aucun doute les plus "discutables" du lot :

$$\forall\, x,\ \big(\neg\,(x = \emptyset) \Rightarrow \exists\, z, (z \in x \wedge z \cap x = \emptyset)\big) \qquad \text{(axiome de fondation)}, \tag{7.47}$$

$$\exists\, x,\ \big(\ \emptyset \in x\ \wedge\ \forall\, z,\ (z \in x \Rightarrow z \cup \{z\} \in x)\ \big) \qquad \text{(axiome de l'infini)}. \tag{7.48}$$

Là où l'axiome de fondation prévient l'existence de chaînes d'appartenances "sans fond" du type "$x \in x$", l'axiome de l'infini met à portée du langage un infini *actuel*, en postulant l'existence d'un ensemble qui ne peut se construire par une succession finie d'opérations élémentaires. C'est, on le verra au chapitre suivant, l'axiome qui permet de parler de l'*ensemble* infini des entiers naturels – on ne se contente donc pas d'un infini *potentiel*.

Récapitulons maintenant la collection des postulats de la *théorie des ensembles* classique, réunis dans le jeu d'axiomes dit de *Zermelo-Fraenkel* :

Axiome d'extensionnalité : Un ensemble est uniquement déterminé par ses éléments.

Axiome de fondation : Il n'existe pas de chaîne d'appartenances récursive.

Axiome de l'infini : Il existe un ensemble infini.

Axiome de la paire : Si a et b sont deux ensembles, on peut construire l'ensemble

$$p = \{\, a, b \,\}.$$

Axiome de la réunion : Si a est un ensemble, on peut construire l'ensemble

$$u \quad = \quad \bigcup a \quad = \quad x_1 \cup x_2 \cup \ldots \quad \text{pour } a = \{x_1, x_2, \ldots\}.$$

Axiome des parties : Si a est un ensemble, on peut construire l'ensemble

$$p = \mathcal{P}(a) = \{\ x \mid x \subseteq a\ \}.$$

Axiomes de remplacement : Si d est un ensemble, si $\varphi(x, y, v_1, \ldots, v_n)$ est une propriété fonctionnelle à n paramètres, alors pour tout choix de ces derniers, on peut construire l'ensemble

$$\begin{aligned} z &= \Big\{\ y \mid \exists\, x \in d, \varphi(x, y, v_1, \ldots, v_n) \text{ vraie } \Big\} \\ &= \Big\{\ \varphi_{(v_1,\ldots,v_n)}(x), \text{ pour } x \text{ décrivant } d\ \Big\}. \end{aligned}$$

Axiomes de séparation : Si a est un ensemble, si $\varphi(x, v_1, \ldots, v_n)$ est une propriété sur x à n paramètres, alors pour tout choix de ces derniers, on peut construire l'ensemble

$$b = \left\{ x \in a \mid \varphi(x, v_1, \ldots, v_n) \text{ vraie } \right\}.$$

Ce jeu d'axiomes est d'une importance considérable. On démontrera en effet dans le chapitre suivant que toutes les notions mathématiques usuelles – nombres entiers, réels ou complexes ; fonctions, suites, etc. – peuvent être "émulées", *modélisées* dans la théorie des ensembles ZF. La cohérence – équivalente à la *consistance*, l'existence d'un modèle d'après le théorème de complétude – de ZF impliquera donc, en cascade, la cohérence de tous les autres systèmes d'axiomes usuels.

Vérité sémantique et théorème de complétude de Gödel

Ces présentations effectuées, revenons à notre théorie générale des *preuves finies*. Les règles de la section 7.2.1, arbitraires, semblent définir un modèle crédible de démonstration... Mais est-il suffisant, sans lacune ?

Partant d'une notion intuitive d'ensemble, de fonction, de relation, on peut définir le fait d'être une $\mathcal{L}$-structure *vérifiant* les axiomes d'une théorie Ax. Le résultat très fort apporté par le théorème de *complétude* de Gödel est alors le suivant :

Théorème 7.1. *Soit φ une $\mathcal{L}$-formule* close*, un énoncé sans variable libre non quantifiée – typiquement, un théorème qui commence par "pour tout ...". Il y a alors équivalence entre :*

φ est conséquence logique de Ax :
Toute $\mathcal{L}$-structure vérifiant les axiomes de Ax vérifie aussi φ.

φ est prouvable dans Ax :
Il existe une preuve formelle de φ dans Ax, au sens de la définition 7.4.

Autrement dit, nos règles de démonstration sont *complètes* : il n'y a pas de résultat "*conséquence nécessaire*" d'un jeu d'axiomes que l'on ne puisse pas démontrer par utilisation du Modus Ponens et autres axiomes logiques.

Il est fastidieux de définir proprement le terme "une structure vérifie une formule", aussi ne démontrerons-nous pas ici le théorème7.1 : c'est un tâche longue, répétitive et pour être honnête, assez rébarbative. Néanmoins, un petit mot sur la preuve. Si montrer le sens réciproque (*prouvable* implique *conséquence logique/nécessaire*) relève de la simple vérification, le sens direct est plus difficile à aborder : comment construire, écrire une preuve à partir d'une "structure" dont on ne sait rien si ce n'est qu'elle vérifie un jeu d'axiomes abstrait ?

La très grande idée de Gödel a été de comprendre qu'un raisonnement par contraposée était possible : à partir d'un jeu d'axiomes Ax et d'un énoncé φ improuvable dans celui-ci, il est possible de construire, à la main, un modèle qui vérifiera Ax et $\neg\varphi$. La construction de ce modèle à partir de mots sur le langage/alphabet $\mathcal{L}$ est une illustration parfaite des principes de la logique mathématique, une science dont l'objet est l'étude du *texte* mathématique.

In-décidabilité, choix d'un système d'axiomes

On l'a vu plus haut : le théorème de complétude de Gödel établit l'équivalence entre *nécessité logique* et *prouvabilité*. Cela signifie-t-il que toute proposition mathématique "vraie" est accessible à la démonstration ? Que tout énoncé est soit "vrai", soit "faux", de négation "vraie".

On serait tenté de le croire... Mais il existe une troisième voie.

Revoir l'intro.

Ni démontrable, ni faux : le paradoxe de l'indécidabilité

Pour le comprendre, nul besoin de recourir à une théorie mathématique complexe. Prenons, tout simplement, un jeu d'axiomes Ax_{ens} formulé dans le langage $\mathcal{L}_{\text{éducation nationale}}$:

1. Dans toute salle de classe, il y a un professeur.
2. Tout professeur est compétent dans la matière qu'il enseigne.

On arrive simplement à attribuer une valeur de vérité – dans la théorie Ax_{ens} – aux énoncés suivants :

$$P \ : \ \text{« Dans toute classe de chinois, quelqu'un parle chinois. »} \qquad \text{« Vrai »}, \tag{7.49}$$

$$Q \ : \ \text{« Il existe une classe d'arabe où personne ne parle arabe. »} \qquad \text{« Faux »}. \tag{7.50}$$

Par Modus Ponens appliqué aux axiomes 1 et 2, Ax_{ens} démontre en effet P et $\neg Q$. Mais que dire alors de ce troisième énoncé :

$$R \ : \ \text{« Dans toute classe de chinois, quelqu'un parle arabe. »} \qquad \text{« ? »}. \tag{7.51}$$

Impossible de dire que (7.51) est un énoncé mal formulé, et pourtant, impossible aussi d'en trouver une démonstration ou une infirmation à partir de Ax_{ens}. Et pour cause : il n'est guère difficile de trouver un modèle de Ax_{ens} qui vérifie R – disons, un bon lycée du Caire –, et un autre modèle qui satisfait tout aussi bien aux exigences de Ax_{ens}, tout en vérifiant $\neg R$ – un collège à Kyoto ! Dans la théorie Ax_{ens}, R est un énoncé qui n'est ni vrai, ni faux – sous réserve de cohérence. Il est *indécidable*.

Les théorèmes d'incomplétude de Gödel

Ce dernier exemple n'est pour tout dire pas très étonnant : notre jeu d'axiomes Ax_{ens}, réduit à deux petits énoncés, n'était pas le plus *complet* qui soit, et la proposition R se trouvait simplement en dehors de son *domaine d'expression*. Non ; la véritable surprise vient maintenant :

Théorème 7.2 (Premier théorème d'incomplétude de Gödel)**.**
Soit Ax une théorie cohérente *et* récursive *– i.e. ses axiomes peuvent être écrits par un programme déterministe – contenant ZF, voire ZF moins l'axiome de fondation.*
Alors Ax est incomplète : il existe des énoncés qui lui sont indécidables*.*

Théorème 7.3 (Second théorème d'incomplétude de Gödel)**.**
Soit Ax une théorie cohérente *et* récursive *contenant l'arithmétique – au sens de Peano, comme expliqué au chapitre suivant.*
Alors la cohérence de Ax est un énoncé indécidable*.*

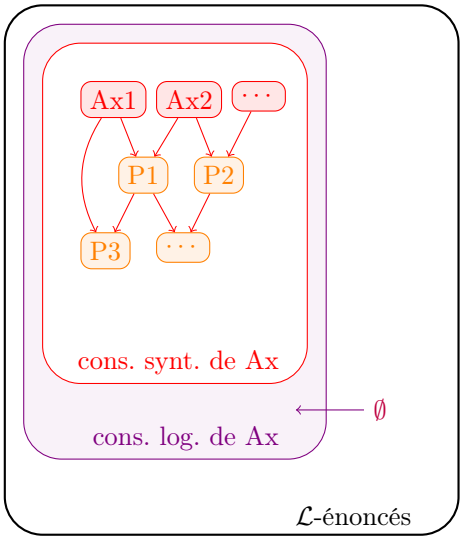
Récursivité et genèse de l'informatique Avant d'aller plus loin, il faut nous pencher sur les énoncés de ces théorèmes ; les décortiquer, comprendre la finesse de leurs hypothèses. L'une d'entre elles, la *récursivité* du système d'axiomes, est d'une importance capitale. En effet, dans un idéal mathématique, il est possible d'imaginer une théorie contenant ZF ou l'arithmétique de Peano qui soit à la fois *cohérente* et *complète* : partons par exemple des axiomes de ZF sur $\mathcal{L}_{ens}$, que l'on suppose cohérents – c'est très raisonnable.

Si ZF est complète, tant mieux : pas besoin de chercher plus loin ! Sinon, *trouvons* une formule φ qui soit indécidable dans ZF : ZF ne prouve ni φ, ni $\neg\varphi$. C'est donc que $\mathrm{ZF} \cup \{\varphi\}$, $\mathrm{ZF} \cup \{\neg\varphi\}$ sont toutes deux cohérentes : on peut alors choisir, à notre convenance, l'une des ces deux théories "enrichies" comme nouveau système d'axiomes de référence. S'il est complet, tant mieux : pas besoin de chercher plus loin ! Sinon... *On recommence.* On devrait alors arriver à une théorie limite – via le "lemme de Zorn", un point technique qu'il serait fort long de détailler – à la fois *cohérente* et *complète*.

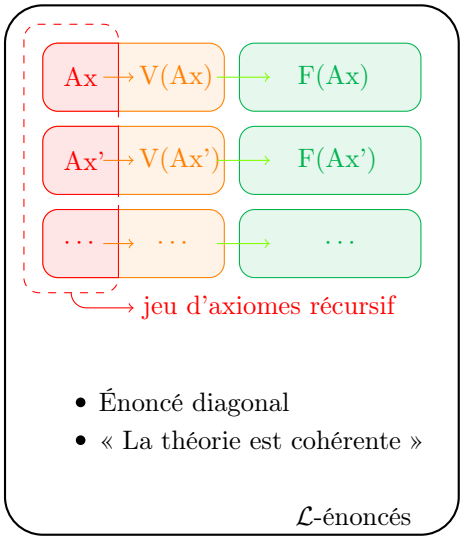
Cette idée de récurrence "transfinie" est fort naturelle... Mais pas très utile en pratique ! En effet, comment *trouver* une formule φ qui soit indécidable dans notre théorie ? Et jusqu'à quand va-t-on *recommencer* comme préconisé ? C'est en cherchant à préciser ces deux points que les logiciens des années 30 ont peu à peu mis à jour une notion claire de *calculabilité*, ou *récursivité* : Une liste de formules, de mots, de nombres sera *récursive* s'il est possible de l'obtenir comme "sortie" d'un programme fini idéalisé, une notion proprement définie.

L'indécidabilité a en fait un analogue dans le monde des programmes. Contrairement à ce que l'on pourrait penser, un programme binaire tournant sur les entiers n'est en effet pas réduit à deux manières de terminer, "0" ou "1" ; il peut aussi... ne jamais s'arrêter. Si l'idée vous intéresse, vous pourrez facilement trouver des renseignement précis au sujet de ce "problème de l'arrêt" : c'est la clé de la preuve des théorèmes d'incomplétude.

Entamé par des mathématiciens soucieux de démontrer la cohérence de leurs théories, cet effort de *formalisation* du raisonnement aura la postérité que l'on connaît : L'acte de naissance de l'informatique est communément daté de 1936, année où Alonzo Church montra dans sa thèse que *récursivité* fonctionnelle et *programmabilité* au sens de Turing étaient équivalentes, formant ensemble le modèle universel de la *calculabilité*.



(a) À partir d'un jeu d'axiomes, on peut a priori définir deux notions de vérité : par les preuves syntaxiques, et par les modèles. Le théorème de complétude de Gödel affirme qu'elles coïncident.



(b) Si un jeu d'axiomes est *cohérent* (l'ensemble des formules vraies et fausses ne s'intersectent pas) et récursif, alors il est incomplet : il existe des énoncés qui ne sont ni vrais, ni faux, au sens de la théorie.

FIGURE 7.1 – Schéma synthétique illustrant les principaux résultats du chapitre. (a) Avec son théorème de complétude, Gödel démontre que tout énoncé *conséquence logique* d'un jeu d'axiomes admet une preuve *syntaxique* à partir de ceux-ci, en utilisant les règles de déduction standards. (b) Si Ax est un jeu d'axiomes, on peut donc définir sans ambiguïtés V(Ax), ensembles des énoncés *conséquences* de Ax, et F(Ax), ensemble des énoncés dont les négations sont dans V(Ax). Si Ax est un jeu d'axiomes fini, encodant une théorie classique comme ZF, on n'est pas étonné d'apprendre qu'il existe des énoncés indécidables pour la théorie, qui ne sont ni dans V(Ax), ni dans F(Ax). On peut donc imaginer d'ajouter de nouveaux axiomes Ax', etc. à la théorie, piochés au fur et à mesure dans l'ensemble des énoncés indécidables. Surprise : si cette procédure de choix est *récursive*, programmable, alors elle ne peut aboutir à une théorie *complète*. Il restera toujours des énoncés bien formés, mais dont la valeur de vérité n'est pas déterminée par la théorie.

Revoir l'intro.

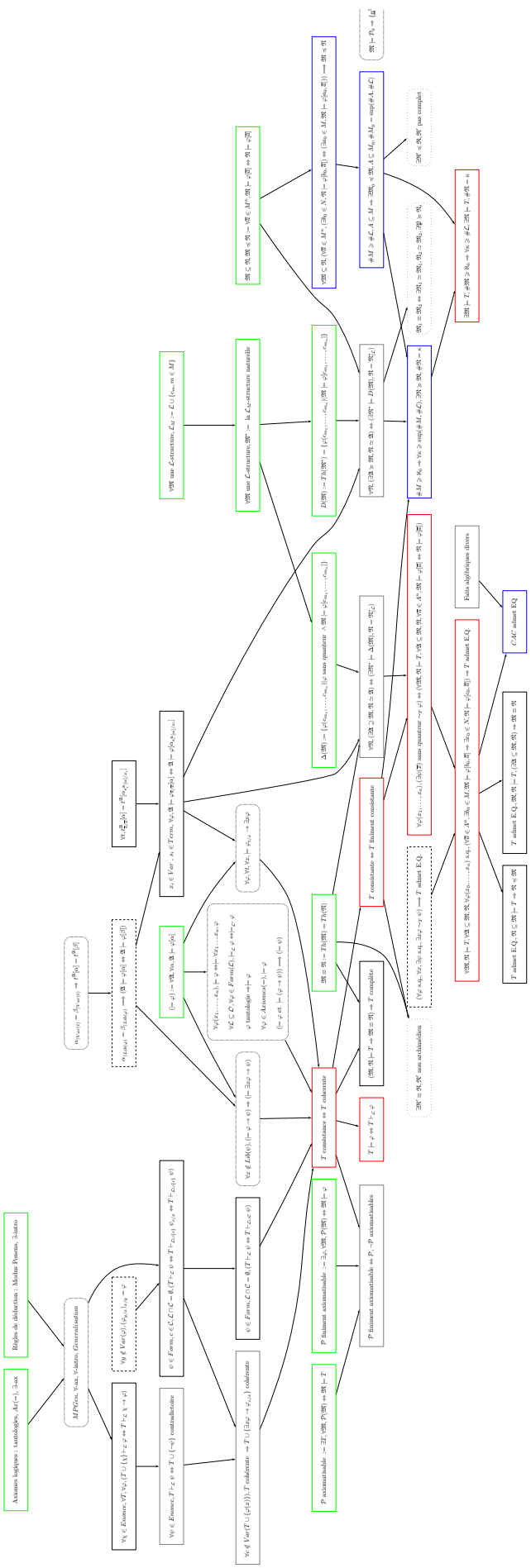


FIGURE 7.2 – Arbre des propositions donnant le plan du cours de *Logique* donné par Martin Hils à l'ENS en 2012-2013. Dans cette première section, on trouve le théorème de complétude de Gödel (à gauche) et ses premières conséquences dont le lemme de compacité, les résultats d'existence en analyse non standard...

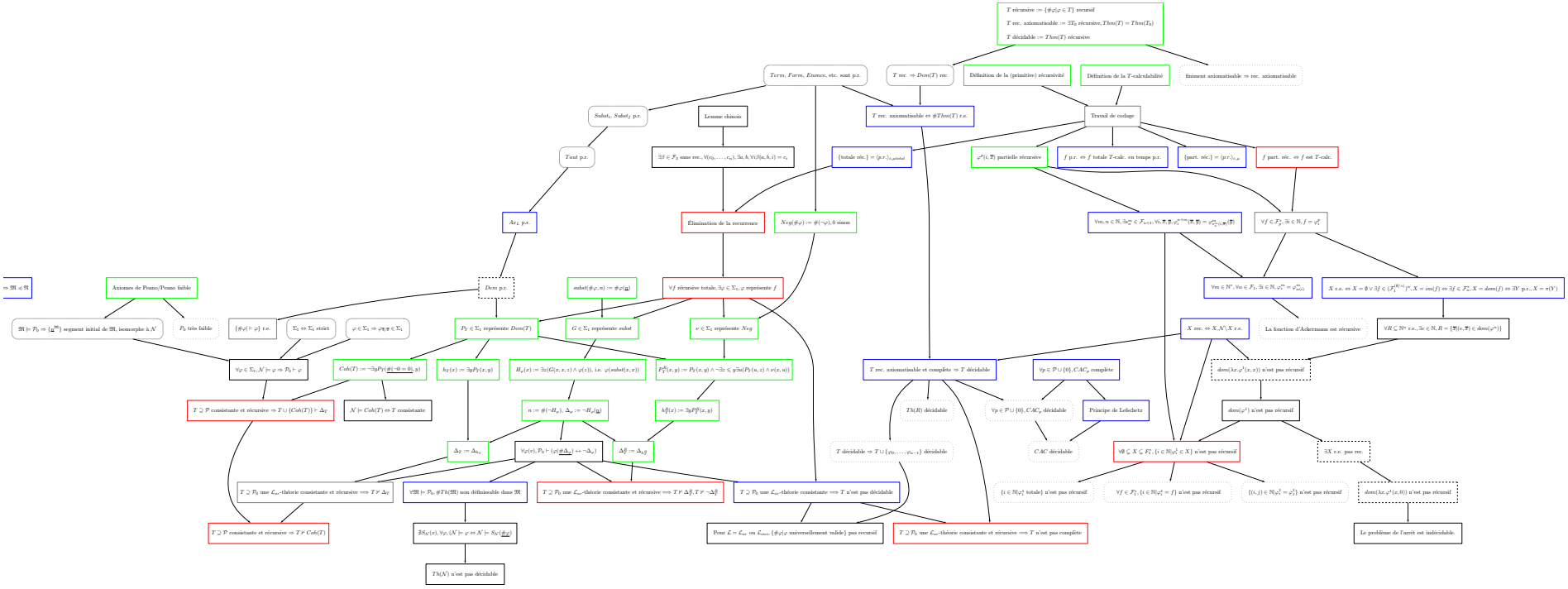


FIGURE 7.3 – La deuxième partie du cours est consacrée à l'encodage numérique des raisonnements syntaxiques (gödelisation, à gauche), à la récursivité (à droite) et au premier théorème d'incomplétude de Gödel, qui tire parti de ces deux approches. Figures réalisées à l'aide du logiciel *Graphviz*.

Importance épistémologique des théorèmes de Gödel Lors de leur publication en 1931, ces théorèmes firent l'effet d'un coup de tonnerre dans la communauté mathématique. Si on était conscient de l'existence d'une troisième éventualité entre le "vrai" et le "faux", on ne doutait pas que tout énoncé mathématique "pertinent" serait à la portée d'un système d'axiomes bien choisis. Mais ce bel optimisme appartient à un temps révolu : d'après le premier théorème d'incomplétude, aucun jeu d'axiome "raisonnable" – c'est à dire *cohérent*, *récursif*, et suffisamment riche pour inclure la théorie des ensembles – ne peut être complet.

Pire : d'après le second théorème d'incomplétude, aucune théorie mathématique *cohérente* et *récursive* capable de modéliser les nombres entiers ne pourra démontrer sa propre cohérence ! L'indécidabilité, loin d'être une bizarrerie sans conséquences, touche donc les fondements mêmes de toute théorie mathématique pertinente.

Tout est ensemble (?)

En 1940, quel bilan tirer du demi-siècle d'efforts épuisées à bâtir les fondements de mathématiques saines ? Nous avons appris à donner un sens précis au terme de "démonstration", et obtenu des moyens *algorithmiques* de vérifier la rigueur d'une preuve. D'une importance cruciale, la vérification automatique de programmes informatiques sensibles – aviation, sécurité, défense – est aujourd'hui un secteur en plein essor.

Mais en "contrepartie", nous avons aussi été mis face aux limites de l'approche axiomatique : le système d'axiomes parfait, à la fois riche, cohérent et complet n'existe pas. Les mathématiciens continueront donc à travailler avec les systèmes d'axiomes de leur choix, adaptés à un domaine spécifique – arithmétique, algèbre, géométrie, probabilités, etc. Pour en assurer la cohérence, ils se contenteront de faire reposer le fardeau sur la théorie des ensembles et son jeu d'axiomes éminemment "crédible", ZF. C'est ce jeu de modélisation, de construction des nombres à partir d'ensembles que je vous proposerai de découvrir au chapitre 8, dès la semaine prochaine.

Références

Parmi les nombreux cours, livres et polycopiés disponibles sur la logique mathématique – généralement abordée par les matheux en fin de licence –, je me suis fortement reposé sur les notes du cours de *Logique* donné en 2012 par Martin Hils au département de mathématiques. Vous pourrez les trouver à l'adresse suivante : `http://www.logique.jussieu.fr/~hils/enseignement/Notes_Cours2012-13.pdf`.

Pour aller plus loin, je vous conseille les premiers chapitres du cours de *Logique et théorie des ensembles* de Patrick Dehornoy (donné dans ce même département en 2006) : avec une introduction moins technique et plus historique que ses successeurs, il me semble plus accessible. Vous retrouverez au chapitre 1 une discussion détaillée sur la nature des ensembles et leur construction actuelle. Vous pourrez le trouver à l'adresse suivante : `http://www.math.unicaen.fr/~dehornoy/surveys.html`.

Sur l'histoire du concept de *calculabilité*, le manuel de Robert Soare, *The History and Concept of Computability*, me semble une bonne référence. Vous pourrez le trouver à l'adresse suivante : `http://www.people.cs.uchicago.edu/~soare/History/handbook.pdf`

Enfin, si vous cherchez simplement à mettre des visages sur les notions présentées ici, une seule référence : l'excellente bande-dessinée *Logicomix*. On y suit les pas du grand logicien Bertrand Russel – un monsieur vraiment remarquable, soit dit en passant –, qui prit une part active à cette aventure de fondation des mathématiques : vous pourrez donc y rencontrer tous les grands noms de la fin XIX^{e} siècle, leurs convictions, leurs échecs... et leurs réussites !

Chapitre 8

Construction classique des ensembles de nombres : $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ et $\mathbb{R}$

Séances 10 et 11

Depuis la plus haute antiquité, les savants, astronomes et marchands utilisent les nombres pour compter. Quoi de plus naturel ? Deux et deux font quatre et quatre font huit...

(Mal)heureusement, depuis la découverte des irrationnels attribuée à Pythagore au VI^{e} siècle avant notre ère, puis des nombres relatifs, voire complexes, on sait que les choses ne sont pas si simples. Pire : en 1874, les travaux menés par Cantor sur les infinis semèrent le doute quand au bien fondé de nos calculs les plus élémentaires.

Et si les nombres n'existaient pas ? On l'a vu, la seule façon de répondre à ces inquiétudes de manière scientifique, "phénoménologique", est de poser la question d'un point de vue *axiomatique* : « Nos présupposés sur les nombres débouchent-ils sur des contradictions ? »

Si nos jeux d'axiomes sont *cohérents*, le théorème 7.1 (complétude de Gödel) assurera l'existence d'une structure qui les vérifie : autrement dit, l'existence d'objets qui se comportent à tous points de vue comme des nombres est garantie. Malheureusement, dans le même temps, les théorèmes 7.2 et 7.3 (incomplétude de Gödel) réduisent à néant les espoirs de *démontrer* cette cohérence.

Devant cette impasse, l'approche des mathématiciens est la suivante : Faute de pouvoir démontrer que la théorie des nombres est *absolument* cohérente, ils vont du moins démontrer qu'elle est *aussi cohérente* que la théorie des ensembles – formalisée par le système ZF, section 7.2.2. Il suffira pour cela de réaliser un modèle de la première dans la deuxième, de *construire* des nombres avec des ensembles ; c'est le travail fondateur que je vous propose de découvrir aujourd'hui.

Les entiers naturels : successeur et récurrence

Plus élémentaire des théories des nombres, l'*arithmétique*, ou science des nombres entiers, est évidemment la première sur notre liste de théories à modéliser. Focalisée sur l'étude de l'ensemble $\mathbb{N} = \{0, 1, 2, \ldots\}$ des entiers positifs, des opérations $+$, $\times$ et des relations qui en découlent, c'est une théorie dont les résultats essentiels sont les suivants :

Théorème 8.1 (Division euclidienne)**.** *Soit a, b deux entiers naturels, $0 < b$.*
Alors il existe un unique *couple d'entiers (q, r) – « quotient », « reste » – tel que :*

$$a = b \times q + r, \qquad 0 \leqslant r < b. \tag{8.1}$$

Théorème 8.2 (Théorème fondamental de l'arithmétique)**.** *Soit n un entier naturel non nul.*

Alors il existe une unique collection finie $q_1 < \cdots < q_r$ de nombres premiers, une unique collection d'exposants $n_1, \ldots, n_r > 0$ tels que

$$n = q_1^{n_1} \times \cdots \times q_r^{n_r}. \tag{8.2}$$

Ceci permet de définir de manière unique la valuation $\mathcal{P}$-adique $\nu : (p, n) \in \mathcal{P} \times \mathbb{N} \mapsto \nu_p(n)$ telle que

$$\forall n \in \mathbb{N}^\star,\ n \ = \ \prod_{p \in \mathcal{P}} p^{\nu_p(n)} \quad = \quad p_1^{\nu_{p_1}(n)} \times p_2^{\nu_{p_2}(n)} \times p_3^{\nu_{p_3}(n)} \times \ldots. \tag{8.3}$$

À chaque entier n correspond une suite $\nu_\cdot(n) : p \in \mathcal{P} \mapsto \nu_p(n) \in \mathbb{N}$ à support fini, et vice-versa.

Axiomes de Peano

Un homme, Giuseppe Peano, a laissé son nom dans l'Histoire pour avoir concocté en 1889 une axiomatisation réussie de l'arithmétique : tous les énoncés classiques sont en effet conséquence du jeu d'axiomes suivant, formulé sur le langage

$$\mathcal{L}_{ar} = \{0, S, +, \times, <\}, \tag{8.4}$$

où « 0 » est une constante, « + » et « × » des fonctions binaires, « < » une relation binaire, et « S » une fonction 1-aire, représentant moralement l'opération *successeur* qui à un entier n associe l'entier suivant, $n + 1$. On y retrouve donc les quatre fonctions constitutives de l'arithmétique : *addition* et *multiplication*, cohérentes entre elles et avec l'*ordre* naturel ; opération *successeur*, qui ouvre la porte au raisonnement par récurrence. Les axiomes ci-dessous se contentent simplement de formaliser les liens entre ces différentes facettes de l'arithmétique :

$\mathbf{(Succ_1)}$:	$\forall\, n,\ \neg\,(Sn = 0)$	« 0 n'est pas successeur »
$\mathbf{(Succ_2)}$:	$\forall\, n,\ \neg\,(n = 0) \Rightarrow (\exists\, m, Sm = n)$	« tout entier non nul est successeur »
$\mathbf{(Succ_3)}$:	$\forall\, m, \forall\, n,\ Sm = Sn \Rightarrow m = n$	« S est injective »
$\mathbf{(Add_1)}$:	$\forall\, n,\ n + 0 = n$	« 0 est neutre à droite pour + »
$\mathbf{(Add_2)}$:	$\forall\, m, \forall\, n,\ m + Sn = S(m + n)$	« + commute avec S »
$\mathbf{(Mul_1)}$:	$\forall\, n,\ n \times 0 = 0$	« 0 est absorbant à droite pour × »
$\mathbf{(Mul_2)}$:	$\forall\, m, \forall\, n,\ m \times Sn = (m \times n) + m$	« × distribue S à gauche »
$\mathbf{(Ord_1)}$:	$\forall\, m, \forall\, n,\ \big(m < n \Leftrightarrow (\neg\,(m = n) \wedge \exists\, k, k + m = n\big)$	« définition de < »

auxquels il faut ajouter, pour chaque $\mathcal{L}_{ar}$-formule $\varphi = \varphi(n, v_1, \ldots, v_p)$ l'axiome de *récurrence* associé :

$$\mathbf{(Rec\,\varphi)}\ :\ \forall\, v_1, \ldots, v_p, \Big(\varphi(0, v_1, \ldots, v_p) \wedge \forall\, k, \big(\varphi(k, v_1, \ldots, v_p) \Rightarrow \varphi(Sk, v_1, \ldots, v_p)\big)\Big) \\ \Longrightarrow\ \forall\, n,\ \varphi\,(n, v_1, \ldots, v_p).$$

Autrement dit : pour tout choix des paramètres $v_1, \ldots, v_p$, si la propriété $\varphi_{(v_1,\ldots,v_p)}$ est vraie en 0 et vérifie le principe de récurrence, alors elle est vraie en tout entier n.

En partant de ce jeu d'axiomes réduit, un travail technique fastidieux permet alors de démontrer toutes les propriétés attendues de l'ensemble des entiers naturels : $<$ est une relation d'ordre strict, car transitive, antiréflexive et antisymétrique ; existence du minimum de toute partie non vide ; associativité, commutativité de + et × ; distributivité de × sur + ; bonne définition de la relation de divisibilité ; division euclidienne ; décomposition en facteurs premiers, etc.

Construction de Von Neumann

Le jeu d'axiomes de Peano est donc suffisamment expressif : pour "garantir" sa cohérence, il reste à le modéliser dans la théorie des ensembles. Suivons ici la construction classique de John Von Neumann, grand logicien et pionnier de l'informatique, concepteur de l'architecture de tous les ordinateurs modernes avec processeur, mémoire et unité arithmétique.

Opération successeur Tout reposera, vous allez le voir, sur cette définition judicieuse de l'opération successeur abstraite σ :

$$\sigma(x,y) \ : \ \forall\ z,\ z \in y \Leftrightarrow (z = x \vee z \in x), \tag{8.5}$$

autrement dit, σ est la formule *fonctionnelle* qui à un ensemble x associe l'ensemble

$$\sigma(x) = x \cup \{x\}, \tag{8.6}$$

bien défini par l'axiome de la paire et de l'union.

Zéro, entiers naturels On prendra pour modèle de 0 l'unique *ensemble vide* vérifiant

$$\emptyset(x) \ : \ \forall\ z, \neg\,(z \in x). \tag{8.7}$$

On peut alors calculer aisément les successeurs de 0, qui représenteront nos nombres entiers naturels :

$$\begin{aligned}
0 &= \{\ \} && (8.8)\\
1 &= S(0) && (8.9)\\
&= \{\ \} \cup \{\{\ \}\} \qquad = \{\{\ \}\} && (8.10)\\
&= \{0\}, && (8.11)\\
2 &= S(1) && (8.12)\\
&= 1 \cup \{1\} \qquad = \{0\} \cup \{1\} && (8.13)\\
&= \{0,1\}, && (8.14)\\
3 &= S(2) && (8.15)\\
&= 2 \cup \{2\} \qquad = \{0,1\} \cup \{2\} && (8.16)\\
&= \{0,1,2\}, && (8.17)\\
n+1 &= \{0,1,2,\ldots,n\}. && (8.18)
\end{aligned}$$

Ordre naturel On le comprend, au vu de (8.18), l'ordre $<$ ne sera pas difficile à définir : il est donné par l'appartenance !

$$< : \ \forall\ x, \forall\ y,\ x < y \Leftrightarrow x \in y. \tag{8.19}$$

Ensemble $\mathbb{N}$ **contenant les entiers naturels** L'ensemble des entiers naturels sera alors, sans surprise, celui qui contient 0 et tous ses successeurs. Attention, l'existence d'une telle "boîte" infinie ne va pas de soi : ce n'est pas parce que l'*on peut exprimer, parler de* tout entier n – infini potentiel, en puissance dans l'infinité de formules du langage – que l'on peut nécessairement représenter par une variable *l'ensemble de tous les entiers*. Pour l'assurer, nous devons recourir à l'*axiome de l'infini*, qui donne l'existence d'un ensemble E tel que :

$$0 \in E \quad \wedge \quad \forall x \in E, \sigma(x) \in E. \tag{8.20}$$

Comme E peut a priori contenir d'autres éléments que les seuls successeurs de 0, on choisit pour ensemble des entiers naturels la *plus petite partie* de E qui contienne 0 et qui soit close par passage au successeur :

$$\text{Clos} = \{\, p \in \mathcal{P}(E) \mid 0 \in p \ \wedge\ \forall x \in p, \sigma(x) \in p \,\} \tag{8.21}$$

$E \in \text{Clos}$, donc Clos est non-vide, ce qui légitime la définition formelle :

$$\mathbb{N} = \bigcap_{p \in \text{Clos}} p, . \tag{8.22}$$

$\mathbb{N}$ est donc lui-même un élément de Clos, contenant 0 et tous ses successeurs sans "surplus".

Opérations algébriques L'addition et la multiplication sont moins évidentes à formaliser... Elles ont néanmoins leurs équivalents ensemblistes, l'*union disjointe* et le *produit cartésien*, respectivement définis par

$$A \uplus B = \{\, (a,0),\ \text{pour } a \text{ décrivant } A \,\} \cup \{\, (b,1),\ \text{pour } b \text{ décrivant } B \,\}, \tag{8.23}$$

$$A \times B = \{\, (a,b),\ \text{pour } a \text{ décrivant } A \text{ et } b \text{ décrivant } B \,\}, \tag{8.24}$$

pour A et B deux ensembles quelconques – le bien-fondé de ces deux opérations est conséquence des axiomes de ZF, comme montré dans l'exercice 7.8. On peut par exemple calculer :

$$\begin{aligned} 2 \uplus 3 &= \{\,0,1\,\} \uplus \{\,0,1,2\,\} && (8.25)\\ &= \{\,(0,0),\,(1,0)\,\} \cup \{\,(0,1),\,(1,1),\,(2,1)\,\} && (8.26)\\ &= \{\,(0,0),\,(1,0),\,(0,1),\,(1,1),\,(2,1)\,\} && (8.27)\\ 2 \times 3 &= \{\,(0,0),\,(0,1),\,(0,2), && (8.28)\\ &\qquad (1,0),\ (1,1),(1,2)\,\} \end{aligned}$$

Puis définir les relations associées :

$$\text{Add}\,(p,a)\ :\ \exists\, x,y,\ p=(x,y)\ \ \wedge\ \ a \simeq x \uplus y, \tag{8.29}$$

$$\text{Mul}\,(p,m)\ :\ \exists\, x,y,\ p=(x,y)\ \ \wedge\ \ m \simeq x \times y, \tag{8.30}$$

où la relation d'équipotence « $\simeq$ », définie par

$$\begin{aligned} \simeq :\ & \forall\, X,Y,\ X \simeq Y && (8.31)\\ & \Leftrightarrow \exists\ \underbrace{f \in Y^X}_{f:X\to Y},\ \underbrace{(\forall\, y \in Y, \exists\, x \in X, f(x)=y)}_{f \text{ surjective, } |X| \geqslant |Y|}\ \wedge\ \underbrace{(\forall\, x_1,x_2 \in X, f(x_1)=f(x_2) \Rightarrow x_1=x_2)}_{f \text{ injective, } |X| \leqslant |Y|} \end{aligned}$$

formalise le fait que X et Y ont le même nombre d'éléments.

Bonne définition des applications du modèle Par construction, $\sigma(x,y)$ est une relation fonctionnelle sur $\mathbb{N}$ à valeurs dans $\mathbb{N}$:

$$\underbrace{\forall\, x \in \mathbb{N},}_{\text{Tout entier } x \text{ a une image entière } y}\ \underbrace{\exists\, y \in \mathbb{N},\ \sigma(x,y)}\ \ \wedge\ \underbrace{\forall\, y_1,y_2 \in \mathbb{N},\ \big(\sigma(x,y_1) \wedge \sigma(x,y_2)\big) \Rightarrow y_1=y_2}_{\text{et celle-ci est unique}}. \tag{8.32}$$

L'application $S : \mathbb{N} \to \mathbb{N}$ qui à n associe $n \cup \{n\}$ est donc bien définie, tout comme la relation $<$ sur $\mathbb{N}\ \times \mathbb{N}$. Montrer que Add et Mul définissent elles aussi des applications $+$ et $\times$ de $\mathbb{N} \times \mathbb{N}$ dans $\mathbb{N}$ est par contre beaucoup plus compliqué : nous l'admettrons ici, mais le lecteur intéressé pourra trouver la démonstration, technique, dans les références indiquées à la fin du chapitre.

Récapitulatif : le modèle de Von Neumann des entiers naturels Le modèle ensembliste standard du langage $\mathcal{L}_{ar}$ est donc donné par :

L'ensemble de base $\mathbb{N}$ défini à partir de l'*axiome de l'infini* équation (8.22),

La relation d'ordre $<$ donnée par l'appartenance $\in$,

L'opération successeur S donnée par la relation $S(n) = n \cup \{ n \}$,

L'addition $+$ donnée par la relation $m + n \simeq m \uplus n$,

La multiplication $\times$ donnée par la relation $m \times n \simeq m \times n$ – au sens du produit cartésien.

On va maintenant démontrer que cette structure ensembliste vérifie les axiomes ($\mathrm{Succ}_{1,2,3}$) et ($\mathrm{Rec}\,\varphi$) pour toute formule φ ; les axiomes de l'ordre, de l'addition et de la multiplication sont en revanche laissés en exercice pour les plus motivés.

($\mathrm{Rec}\,\varphi$) : Axiomes de récurrence Fixons $\varphi(n, v_1, \ldots, v_p)$ une formule, et $v_1, \ldots, v_p$ des paramètres quelconques. On suppose que $\varphi(0, v_1, \ldots, v_p)$ est vérifiée, et que pour tout k, $\varphi(k, v_1, \ldots, v_p) \Rightarrow \varphi(Sk, v_1, \ldots, v_p)$. D'après l'axiome de *séparation*, on peut considérer l'ensemble

$$P = \{ n \in \mathbb{N} \mid \varphi(n, v_1, \ldots, v_p) \text{ est vraie} \} \tag{8.33}$$

des entiers qui vérifient $\varphi(\,\cdot\,, v_1, \ldots, v_p)$. On a par définition $P \subseteq \mathbb{N}$, et nos hypothèses sur φ assurent que $P \in \mathrm{Clos}$ comme défini équation 8.21 : on a alors $\mathbb{N} \subseteq P$, i.e. $\forall n \in \mathbb{N}, \varphi(n, v_1, \ldots, v_p)$.

(Succ_1) : 0 n'est pas successeur Supposons en effet disposer d'un entier n tel que

$$n \cup \{ n \} = Sn = 0 = \{\ \}. \tag{8.34}$$

On aurait alors $n \in \{\ \}$, ce qui contredit la définition de l'ensemble vide.

(Succ_2) : Tout entier non nul est successeur C'est une conséquence directe de l'axiome de récurrence ($\mathrm{Rec}\,\psi$) associé à la formule

$$\psi(n) \ : \ (n = 0) \quad \vee \quad \exists\, k,\ Sk = n, \tag{8.35}$$

démontré plus haut. On a en effet $\psi(0)$ et, pour tout k, $\psi(Sk)$, donc en particulier $\psi(k) \Rightarrow \psi(Sk)$.

(Succ_3) : S est injective Prenons m, n dans $\mathbb{N}$ tels que $Sm = Sn$, et supposons par l'absurde que $\neg\,(m = n)$. Comme

$$m \cup \{ m \} = n \cup \{ n \}, \tag{8.36}$$

on a $m \in n \cup \{ n \}$, puis $m \in n$. De même, on a $n \in m \cup \{ m \}$, qui implique $n \in m$: combinant les deux, on a $n \in m \in n$, ce qui contredit l'axiome de *fondation*. C'est donc que $n = m$.

Entiers relatifs et nombres rationnels : premières structures algébriques

Soupir de soulagement : L'existence des entiers naturels est maintenant "assurée". On peut donc *compter* sans risquer de s'en mordre les doigts !

Les entiers relatifs facilitent la mise en équations

Mais que penser des entiers relatifs, les fameux "nombres négatifs" ? Les mathématiciens déconseillèrent longtemps l'usage de ces outils algébriques, soi-disant "absurdes"... Et pourtant, au XVIII^{e} siècle, ces "faux" nombres – pour reprendre la terminologie cartésienne – s'étaient partout imposé.

Pour comprendre ce succès, il faut mettre le doigt sur les lacunes d'une arithmétique des entiers naturels pourtant bien dotée. Une addition compatibles avec l'ordre et commutative, une multiplication qui la distribue... $(\mathbb{N}, +, \times, <)$ a presque tout, et même un *neutre* pour la loi « + », le fameux nombre *zéro* qui vérifie

$$\forall\ n \in \mathbb{N},\ 0 + n = n = n + 0. \tag{8.37}$$

Formalisation des problèmes linéaires Mais si les entiers naturels permettent de *compter*, il se révèlent malcommodes dès qu'il s'agit de *comptabiliser* : Pour résoudre des problèmes pratiques, le grand savant persan Al-Khwârizmî (al-gorithme) est ainsi contraint d'écrire un traité entier, l'*Abrégé du calcul par la restauration et la comparaison*, pour décrire – sans symboles – les méthodes de résolutions de problèmes *linéaires*, puis *quadratiques*. L'*al-jabr* – ou opération de *réduction* – sert alors à ramener un problème donné à un problème connu de coefficients positifs, tels que :

$$(P_{a,b,d}) :\ \text{trouver x tel que}\ \ a \times x + b = d, \tag{8.38}$$

$$(Q_{a,b,d}) :\ \text{trouver x tel que}\ \ a \times x + b = c \times x, \tag{8.39}$$

$$(R_{a,b,c,d}) :\ \text{trouver x tel que}\ \ a \times x + b = c \times x + d, \tag{8.40}$$

qui sont alors vus comme des problèmes distincts, nécessitant chacun une méthode de résolution ad hoc. L'unification, la simplification conceptuelle au travers de notations symboliques abstraites n'arrivera qu'au tournant du XVII^{e} siècle, avec entre autres les travaux de Viète (1540–1603) puis Descartes (1596–1650).

Le premier pas vers la résolution *efficace* de tels problèmes est de mettre ceux-ci sous une forme *canonique*, de "faire passer à gauche le membre de droite". Ce qui justifiera une telle manœuvre, c'est l'existence d'*opposés* ; autrement dit, l'existence pour tout $n \in \mathbb{N}$ d'un nombre n' tel que

$$n + n' = 0 = n' + n. \tag{8.41}$$

Notez que ces *opposés* des entiers naturels sont nécessairement uniques : si n' et n'' sont deux opposés du même entier n, alors

$$0 = n' + n \Longrightarrow 0 + n'' = (n' + n) + n'' \Longrightarrow n'' = n' + 0 = n'. \tag{8.42}$$

L'unique opposé de n, a priori formel, sera le *nombre négatif* « $-n$ ». Grâce à lui, on pourra réduire l'étude des trois classes de problèmes P, Q, R à celle de la simple équation bi-paramétrée

$$(E_{a,b}) :\ \text{trouver x tel que}\ \ a \times x + b = 0. \tag{8.43}$$

On a en effet – ajouter l'opposé du membre de droite aux deux côtés de l'équation :

$$\begin{aligned}
a\times x+b=d &\quad\Leftrightarrow\quad a\times x+(b-d)=0 &&\text{i.e.}\quad (P_{a,b,d}) \Leftrightarrow (E_{a,b-d}), &(8.44)\\
a\times x+b=c\times x &\quad\Leftrightarrow\quad (a-c)\times x+\ b\ =0 &&\text{i.e.}\quad (Q_{a,b,c}) \Leftrightarrow (E_{a-c,b}), &(8.45)\\
a\times x+b=c\times x+d &\quad\Leftrightarrow\quad (a-c)\times x+(b-d)=0 &&\text{i.e.}\quad (R_{a,b,c,d}) \Leftrightarrow (E_{a-c,b-d}). &(8.46)
\end{aligned}$$

Pour garantir le bien-fondé des opérations ci-dessus, il ne reste maintenant plus qu'à démontrer que "l'axiome de l'opposition" n'entame pas la cohérence de l'arithmétique...

Modélisation dans la théorie ZF Démontrer l'existence *ensembliste* de $\mathbb{Z}$ ne présente aucune difficulté : on conviendra simplement que

$$\begin{aligned}
\mathbb{Z} &= \{\,0\,\}\times(\mathbb{N}\backslash\{\,0\,\})\ \ \cup\ \{\,1\,\}\times\mathbb{N} &(8.47)\\
&= \underbrace{\{\ (0,n),\ \text{pour } n \text{ entier non nul}\ \}}_{(0,n)\text{ représente } -n}\ \cup\ \underbrace{\{\ (1,n),\ \text{pour } n \text{ entier}\ \}}_{(1,n)\text{ représente } +n}. &(8.48)
\end{aligned}$$

L'ordre naturel $<$ sera étendu tout aussi simplement en décrétant que

$$\begin{aligned}
\forall\ (s,m),(t,n)\in\mathbb{Z},\ (s,m)<(t,n) \Leftrightarrow\ \Big(&\ (s=0)\wedge(t=1) && \text{« } -m<+n \text{ »} &(8.49)\\
&\vee\ (s=t=0)\wedge(n<m) && \text{« } -m<-n \text{ »} &(8.50)\\
&\vee\ (s=t=1)\wedge(m<n)\ \Big). && \text{« } +m<+n \text{ »} &(8.51)
\end{aligned}$$

« Que devins-je quand je m'aperçus que personne ne pouvait m'expliquer comment il se faisait que : moins par moins donne plus ? » Si le professeur du petit Henri Beyle (aka. Stendhal) avait été plus instruit, il aurait pu justifier la fameuse règle des signes « $-\times-=+$, $-\times+=-$ » de manière fort simple : c'est la seule règle *cohérente* avec la distributivité de la multiplication sur l'addition.

Conséquence de l'axiome (Mul_2) de l'arithmétique de Peano, cette loi stipule en effet que

$$\forall\ m,n,p,\ m\times(n+p)=m\times n+m\times p. \tag{8.52}$$

Pour $+m$, $+n$ fixés dans $\mathbb{N}$, on veut donc avoir

$$\begin{aligned}
0=m\times 0 &= m\times((-n)+n) &(8.53)\\
&= m\times(-n)+m\times n &(8.54)\\
\text{puis}\quad 0-(m\times n) &= m\times(-n)+m\times n-(m\times n) &(8.55)\\
\text{i.e.}\quad -(m\times n) &= m\times(-n), &(8.56)
\end{aligned}$$

et on trouve de même $-(m\times n)=\ (-m)\times n$: voilà pour « $-\times+=-$ ». Il reste alors à développer le double produit

$$\begin{aligned}
0 &= ((-m)+m)\times((-n)+n) &(8.57)\\
&= (-m)\times(-n)+(-m)\times n+m\times(-n)+(m\times n) &(8.58)\\
&= (-m)\times(-n)-\ (m\times n)\ -\ (m\times n)\ +(m\times n) &(8.59)\\
\text{pour avoir}\quad m\times n &= (-m)\times(-n). &(8.60)
\end{aligned}$$

Cette règle trouvée, il n'est pas difficile de l'implémenter à l'aide de fonctions explicites de $\mathbb{Z}\times\mathbb{Z}$ dans $\mathbb{Z}$, $+_{\mathbb{Z}}$ et $\times_{\mathbb{Z}}$: on aura donc *modélisé* par des ensembles la théorie des entiers relatifs qui s'en trouvera par là-même "assurée".

Les nombres rationnels permettent de résoudre les problèmes linéaires

Ce que l'on a fait pour l'addition, on le fait naturellement pour la multiplication : on construit formellement le *corps* des rationnels en associant à tout nombre entier relatif non nul n un – unique – inverse n^{-1} tel que

$$n^{-1} \times n = 1 = n \times n^{-1}, \tag{8.61}$$

où le *neutre* de la multiplication 1 est l'unique élément tel que

$$\forall\ x,\ 1 \times x = x = x \times 1. \tag{8.62}$$

Notez que l'on ne saurait doter 0 d'un inverse sans mettre en péril la cohérence de notre théorie : pour tout entier n,

$$n \times 0 = 0 \tag{8.63}$$

$$\text{impliquerait} \quad n \times 0 \times 0^{-1} = 0 \times 0^{-1} \tag{8.64}$$

$$\text{i.e.} \qquad n = 1, \qquad \text{ce qui est gênant si } n = 2. \tag{8.65}$$

Intérêt pratique : la résolution d'équations linéaires On a vu plus haut que l'existence d'*opposés* permet de ramener toute équation linéaire non dégénérée sur $\mathbb{Z}$ à un problème de la forme

$$(E_{a,b}) : \ \text{trouver x tel que} \ \ a \times x + b = 0, \tag{8.66}$$

où a et b sont des entiers, a non nul. Malheureusement, dans $\mathbb{Z}$, cette équation n'a pas toujours une solution... Il suffit pour cela que b ne soit pas un *multiple* de a.

L'introduction d'*inverses* permet de lever cette fastidieuse condition de divisibilité : le problème linéaire $E_{a,b}$ devient équivalent à la simple équation solution

$$(S) \ : \ \ x = -\,b \times a^{-1}. \tag{8.67}$$

Autrement dit, travailler dans $\mathbb{Q}$ permet d'associer à tout couple $(a, b) \in (\mathbb{Z} \setminus \{\,0\,\}) \times \mathbb{Z}$ une unique solution $x_{\text{sol}} = -\,b \times a^{-1}$.

Travailler dans le *corps* des rationnels, où tout élément non nul est inversible, nous a donc permis d'unifier sous un même formalisme les équations

$$(F) \ : \ \ 2 \times x + 4 = 0 \qquad \text{et} \qquad (G) \ : \ \ 2 \times x + 5 = 0. \tag{8.68}$$

Un progrès salutaire et nécessaire, vous en conviendrez !

Classes d'équivalences et passage au quotient Comment construire une représentation ensembliste de $\mathbb{Q}$? Contrairement à $\mathbb{Z}$ qui n'est après tout qu'un simple "symétrisé" de $\mathbb{N}$, l'ensemble des rationnels semble nettement plus gros et complexe à définir. Pour y arriver, nous formaliserons la notion de *fraction* à l'aide de *classes d'équivalences*.

Rappelez-vous : au chapitre 1, nous avons vu comment "rendre égales des choses qui ne le sont pas" – les configurations du jeu de taquin. Dans le même esprit, on défini la relation d'*équivalence* « $\sim$ » sur les couples "numérateur/dénominateur" par

$$\forall\ (n_1, d_1), (n_2, d_2) \in \mathbb{Z} \times (\mathbb{Z} \setminus \{\,0\,\}), \quad (n_1, d_1) \sim (n_2, d_2) \quad \Leftrightarrow \ n_1 \times d_2 = n_2 \times d_1. \tag{8.69}$$

On définit alors la *fraction* $\overline{(n,d)}$, plus couramment notée $\frac{n}{d}$, par

$$\frac{n}{d} = \overline{(n,d)} = \Big\{(a,b) \in \mathbb{Z} \times (\mathbb{Z} \setminus \{0\}) \mid (n,d) \sim (a,b)\Big\}. \tag{8.70}$$

On trouve par exemple que

$$\frac{1}{2} = \{\ldots, (-1,-2), (1,2), (2,4), (3,6), \ldots\} = \frac{2}{4} = \ldots, \tag{8.71}$$

$$\frac{1}{3} = \{\ldots, (-1,-3), (1,3), (2,6), (3,9), \ldots\} = \frac{2}{6} = \ldots, \tag{8.72}$$

et on verra par là qe notre astuce ensembliste a permis de légitimer les "simplifications" de fraction enseignées au collège. On identifie naturellement tout entier n de $\mathbb{Z}$ avec la fraction $\frac{n}{1}$ associée ; prolonger les opérations usuelles que sont l'addition, la multiplication et la comparaison ne pose alors aucune difficulté – disons que c'est un bon exercice –, et l'on constatera alors avec bonheur que, pour tout entier n non nul,

$$\frac{n}{1} \times \frac{1}{n} = \frac{n \times 1}{1 \times n} = \frac{n}{n} = \frac{1}{1}. \tag{8.73}$$

qui est l'élément neutre de la multiplication sur $\mathbb{Q}$, « $1_{\mathbb{Q}}$ » : tout entier et, plus largement, tout rationnel non nul est donc inversible dans $\mathbb{Q}$, comme annoncé.

Les nombres réels ou la puissance du continu

En passant de la théorie des entiers naturels, qui permet de *compter*, à la théorie des nombres rationnels, qui permet de *calculer*, nous avons fait un grand pas. Mais le confort offert par l'ensemble $\mathbb{Q}$ des nombres rationnels est en fait tout relatif : il est, c'est la cas de le dire, plein de lacunes !

Problème des valeurs intermédiaires

Reprenons nos calculs comptables de la section précédente, en récapitulant les étapes permettant de passer d'un problème pratique à une solution explicite :

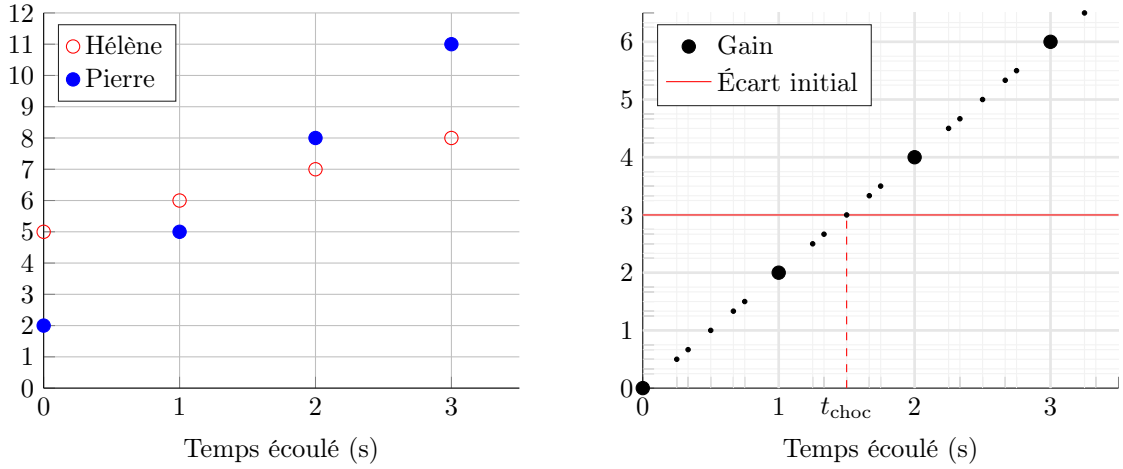
Problème brut Ce matin, Hélène et Pierre font un petit tour. À peine parti, Pierre se voit déjà distancé par Hélène : malgré sa faible vitesse de 1 m/s elle se trouve déjà à 5m de la maison. Pierre, à deux pas seulement de son jardin, donne donc un petit coup de rein pour se porter à la vitesse record de 3 m/s.
Question : Combien de temps mettra-t-il pour rattraper sa compagne ?

Mise en équation Résoudre en t l'équation (R) : $3 \times t + 2 = 1 \times t + 5$.

Mise sous forme canonique On se ramène à (E) : $2 \times t = 3$.

Résolution On trouve l'unique solution $t_{rec} = 3/2$ – qui correspond à une distance au perron de $3t_{rec} + 2 = t_{rec} + 5 = 6,5$m.

On peut en fait interpréter ce programme de résolution comme un algorithme de recherche d'un point d'intersection entre deux droites affines : là où le travail dans $\mathbb{Z}$ a permis de ramener par soustraction l'une des deux droites à une constante, la structure de $\mathbb{Q}$ a permis de garantir l'existence d'un point d'intersection entre la pente $y = 2 \times x$ et la droite horizontale $y = 3$, au point d'abscisse $x = 3/2$ – voir à ce sujet la figure 8.1.



(a) Problème brut, représenté ici en coordonnées entières. On cherche à déterminer l'abscisse du "point d'intersection" entre les deux trajectoires.

(b) Par soustraction, on se ramène à l'étude du *gain* de distance de Pierre sur Hélène. Travailler dans $\mathbb{Q}$ garantit alors l'existence de l'instant t_{choc} auquel Pierre rattrape sa compagne.

FIGURE 8.1 – Illustration graphique de la résolution du problème des promeneurs.

Problèmes quadratiques Les problèmes linéaires sont donc à la portée du calcul rationnel. Mais que penser de situations plus complexes ? La chute d'un corps est un bon exemple : ici, point de promeneur marchant à vitesse constante, mais une bille tombant d'un tabouret (de hauteur h indéterminée) dans une pièce soumise à un champ de gravité constant $g = 2\,\mathrm{m/s^2}$ – pour simplifier ; à Paris, on aurait plutôt $g = 9.81\,\mathrm{m/s^2}$.
Question : Après combien de temps touchera-t-elle le sol ?

On sait depuis Galilée que dans une telle situation, la distance d entre la bille et son point de départ s'accroît suivant la loi quadratique

$$d(t) = \frac{1}{2} g\, t^2. \tag{8.74}$$

On cherche donc à trouver l'abscisse t_{choc} du point d'intersection entre la courbe de chute $d(t) = t^2$ et la hauteur constante h. *Problème* : pour $h = 2\,\mathrm{m}$, aucun instant $t \in \mathbb{Q}$ ne convient.

Irrationalité de $\sqrt{2}$ Un tel instant t_{choc} devrait en effet vérifier

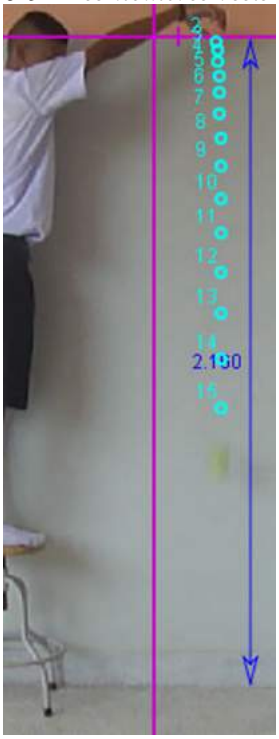
$$t_{\text{choc}}^2 = h = 2, \tag{8.75}$$

ce qui est *impossible* si t_{choc} s'écrit p/q avec p et q deux entiers. On sait en effet – théorème 8.2 – que tout entier n admet une unique "valuation 2-adique" $\nu_2(n)$, qui indique le nombre maximal de fois qu'il est possible de diviser n par 2 avant de tomber sur un nombre impair. On a par exemple $\nu_2(20) = \nu_2(4 \times 5) = 2$ et $\nu_2(56) = \nu_2(8 \times 7) = 3$. Or, supposons par l'absurde disposer de p et q tels que

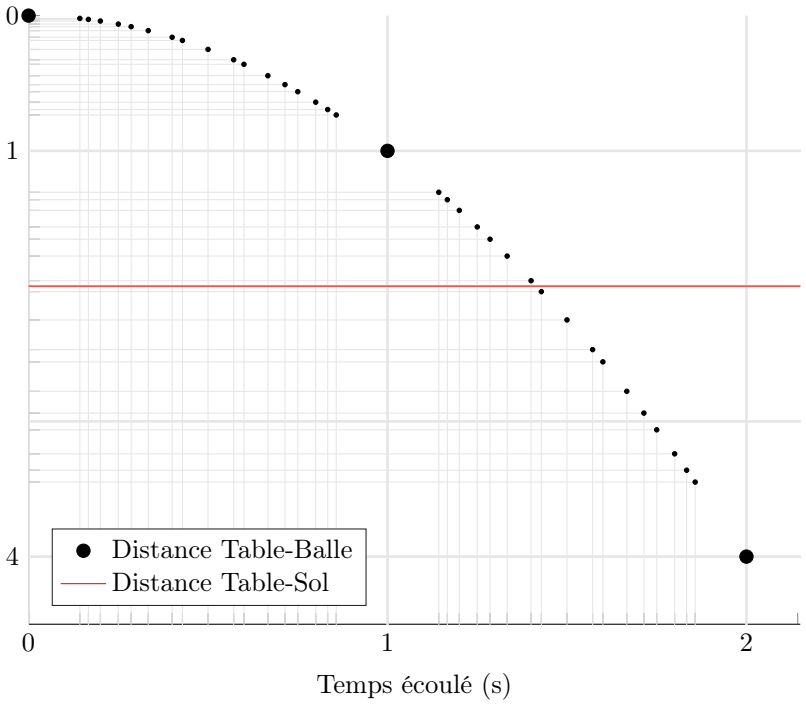
$$p^2/q^2 = 2 \qquad \text{i.e.} \qquad p^2 = 2 \times q^2. \tag{8.76}$$

On aurait alors

$$2 \times \nu_2(p) = \nu_2(p^2) = \nu_2(2 \times q^2) = 1 + 2 \times \nu_2(q), \tag{8.77}$$



(a) Chute libre. Tiré de *Digital video analysis of falling objects in air and liquid using Tracker*, Sirisathitkul et Al.



(b) Tracé de la fonction $d : t \mapsto t^2$ sur $\mathbb{Q}$, tête en bas. On représente ici en abscisse les points rationnels dont le dénominateur est inférieur ou égal à 7. L'axe des distances est criblé d'un ensemble de valeurs "carrées d'un rationnel" qui est *dense*, mais dont 2 n'en fait pas partie.

FIGURE 8.2 – Illustration graphique de l'irrationalité de $\sqrt{2}$: 2 n'a pas d'antécédent dans $\mathbb{Q}$ par la fonction de mise au carré, ce qui semble impliquer que l'instant où la balle entre en contact avec le sol *n'est pas défini* : il est impossible de l'écrire en fractions de seconde.

ce qui est absurde puisque le membre de gauche est pair, et celui de droite, impair ; Cqfd.

Connue depuis l'antiquité, l'irrationalité de $\sqrt{2}$ jette donc une ombre sur le calcul rationnel, incapable d'exprimer autre chose que la linéarité.

Des p'tits trous On a en fait mis en évidence que l'ensemble des rationnels n'est pas un milieu *continu*. D'une fonction *régulière* – sans saut – définie sur un intervalle continu, on attend en effet la propriété des *valeurs intermédiaires* :

Propriété 8.1. (Valeurs intermédiaires) Si f est définie entre a et b, alors toute valeur image comprise entre $f(a)$ et $f(b)$ admet au moins un antécédent par f entre a et b.
Autrement dit – si $f(a) > f(b)$, ou $a > b$, inverser les ordres –,

$$\forall\, y,\ \Big(f(a) \leqslant y \leqslant f(b)\ \Rightarrow\ \exists\, x,\ \big(a \leqslant x \leqslant b\ \wedge\ f(x) = y \big) \Big). \tag{8.78}$$

Comme illustré figure 8.2, $\mathbb{Q}$ est moralement "rempli de trous" – ce qui l'empêche d'accorder un antécédent pour 2 à la fonction $t \mapsto t^2$. Une solution pratique consiste alors à *enrichir* l'ensemble des rationnels par l'ajout de solutions d'équations polynomiales : racines carrées, cubiques, etc. À la limite, on obtient un ensemble bien défini, celui des nombres *algébriques* qu'il est possible d'équiper des opérations usuelles.

Le continu existe-t-il ?

Pendant près de deux millénaires, les mathématiciens se satisfirent de cette rustine sans creuser beaucoup plus loin. Après tout, leurs travaux portaient avant tout sur le calcul de quantités algébriques ou analytiques, et les critères de rigueur n'étaient pas les mêmes qu'aujourd'hui. On pense par exemple à cette rafraîchissante preuve par Cauchy du "théorème" des valeurs intermédiaires : « Comme l'ordonnée constante $[y]$ se trouve comprise entres les coordonnées $[f(a)]$ et $[f(b)]$ des deux points que l'on considère, la droite [horizontale de niveau y] passera nécessairement entre ces deux points, ce qu'elle ne peut faire sans rencontrer dans l'intervalle la courbe [représentative de f] ». L'intuition géométrique était donc suffisante – pour être tout à fait honnête, ajoutons que Cauchy proposera une démonstration "analytique" plus conforme aux canons actuels dans les appendices de son cours d'analyse à l'école polytechnique.

Cantor et la puissance du continu Tout ceci dura jusqu'au jour de décembre 1873 où, brisant le statu quo, un résultat mit en péril tout l'édifice mathématique construit jusqu'alors – motivant ainsi la refondation des mathématiques présentée dans les chapitres 2 et 3. Il s'agit du théorème suivant, (presque) découvert par mégarde :

Théorème 8.3. *(Puissance du continu ; Cantor, 1874)*
L'ensemble de réels $[0,1]$ *n'est pas dénombrable.*
*En d'autres termes, il est impossible d'*énumérer *les points d'un segment continu.*

Démonstration. On raisonne par l'absurde. Supposons disposer d'une énumération $(x_n)_{n\in\mathbb{N}^\star}$ de $[0,1[$, c'est à dire d'une suite telle que

$$[0,1[= \{\, x_1,\, x_2,\, x_3,\, \dots \}. \tag{8.79}$$

Pour chaque indice n, on considère l'unique développement décimal propre de x_n – i.e. celui qui ne se termine pas par $\dots 9999 \dots$ –, que l'on reporte à la n$^{\text{e}}$ ligne d'un tableau infini :

Réel \ Décimale	0	1	2	3	4	$\cdots$
x_1	0,	**0**	2	4	0	$\cdots$
x_2	0,	2	**0**	5	6	$\cdots$
x_3	0,	0	0	**5**	7	$\cdots$
x_4	0,	1	3	4	**8**	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$
y	0,	1	1	0	0	$\cdots$

On construit alors le développement décimal du réel y à partir de la diagonale "de Cantor" du tableau – ici en gras –, en décrétant que la n^{e} décimale de y vaut 1 si le coefficient diagonal associé vaut 0, et 0 sinon. y est bien un réel de $[0,1[$, mais il ne peut, par construction, être inclus dans l'énumération (x_n) : sa k^{e} décimale diffère de celle de x_k, pour tout k.

Ceci est en contradiction avec l'hypothèse de départ, qui est donc absurde ; Cqfd. □

Questions ouvertes et paradoxes Quand on sait qu'il est possible d'énumérer *toutes* les formules du langage – énumérez toutes les formules à une lettre, puis celles à deux lettres, à trois lettres, etc. –, on comprend mieux la portée de ce théorème : l'existence d'un segment continu implique l'existence de nombres *qu'il est impossible de décrire explicitement.*

Ce paradoxe montre en fait la richesse étonnante d'une notion a priori intuitive, celle de *milieu continu.* La question de sa pertinence pour une description mathématique du monde physique est d'ailleurs un problème non résolu à ce jour. Prenez un verre d'eau ; une description

atomiste, granulaire de son contenu donne la garantie qu'à partir d'un état donné, une et une seule évolution peut suivre : c'est le théorème fondamental de la théorie des systèmes dynamiques, que nous développerons section B.4.2.

Avec une description continue par contre – équations de Navier-Stokes –, rien de tout cela n'est sûr : il n'y a a priori pas de *déterminisme* dans l'évolution du système, et pire encore, il n'est même pas garanti qu'il y ait seulement *existence* d'une évolution possible du système « verre d'eau » ! Apporter une réponse à ces questions est aujourd'hui le Graal des physiciens spécialisés en mécanique des fluides, mais rien ne laisse penser qu'il sera trouvé de notre vivant...

Les coupures de Dedekind

Je pense vous avoir convaincu de la difficulté qu'il y a à justifier (méta–)physiquement la notion de *milieu continu*. Sans nous plonger dans de profonds abîmes théoriques, nous nous contenterons donc modestement de montrer que, si la théorie des milieux continus est paradoxale et fortement non-triviale, elle n'en reste pas moins cohérente : il est possible de la modéliser dans la théorie des ensembles.

Axiome de la borne supérieure Avant tout, rappelons que nous essayons de démontrer l'existence d'une structure $(\mathbb{R}, +, \times, <)$ telle que :

— $\mathbb{Q}$ s'identifie naturellement à une partie dense de $\mathbb{R}$,

— addition et produit soient compatibles avec l'ordre,

— tout élément non nul soit opposable et inversible,

— la propriété des valeurs intermédiaires soit vérifiée.

Plutôt que d'axiomatiser la propriété des valeurs intermédiaires – qui nécessite de définir proprement la notion de fonction *continue* –, on préfère généralement assurer une propriété plus élémentaire : l'*axiome de la borne supérieure*, qui affirme que toute partie non vide et majorée admet une *borne supérieure* i.e. un plus petit majorant. Formellement, cette propriété – qui, on le verra au théorème B.2.1, implique la propriété des valeurs intermédiaires – s'écrit

$$\begin{aligned}(\text{Sup}) \ : \ \forall\, A \in \mathcal{P}(\mathbb{R}),\ & \underbrace{(\exists\, x,\, x \in A \wedge \exists\, M, \forall\, a \in A, a \leqslant M)}_{A \text{ non vide et majoré}} \\ \Longrightarrow \exists\, s,\ & \underbrace{(\forall\, a \in A, a \leqslant s)}_{s \text{ majore } A} \ \wedge \ \underbrace{(\forall\, M, \forall\, a \in A, a \leqslant M \Rightarrow s \leqslant M)}_{\text{et il est le plus petit à le faire}}.\end{aligned} \tag{8.80}$$

Surprenant au premier abord, cet énoncé a en fait l'élégance de traduire par de simples comparaisons l'absence de trous, ou *complétude*, de la droite réelle. Regardons par exemple dans $\mathbb{Q}$ le cas de la partie

$$D = \{\, x \in \mathbb{Q} \mid \ x \leqslant 0 \ \vee \ x^2 < 2 \,\}. \tag{8.81}$$

D est non vide et indubitablement majoré – par 3, par exemple – et pourtant, le plus petit majorant de D, noté $\sup D$, ne saurait exister dans $\mathbb{Q}$, comme démontré figure 8.3.

À l'inverse, dans l'idée que l'on se fait de la droite réelle, $\sup D$ n'est pas difficile à trouver : c'est tout simplement $\sqrt{2}$! Filer ce constat nous mène naturellement, en suivant Dedekind, à représenter un réel r par l'ensemble des rationnels qui lui sont strictement inférieur :

$$\text{« } r = \{\, x \in \mathbb{Q} \mid \ x < r \,\} = \mathbb{Q} \cap]-\infty, r\,[\text{ ».} \tag{8.82}$$

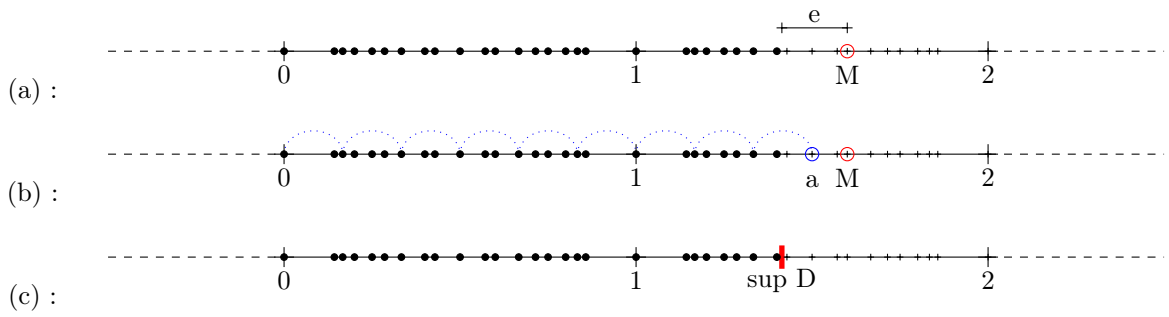


FIGURE 8.3 – Défini équation 8.81, l'ensemble D des rationnels dont le carré est inférieur à 2 – ou qui sont négatifs – n'admet pas de borne supérieure rationnelle. Comme à la figure 8.2, on représente ici les rationnels dont le dénominateur est inférieur ou égal à 7 afin de rendre apparente la granularité de $\mathbb{Q}$.
(a) : Soit M un majorant de D. Comme $x^2 = 2$ n'admet pas de solution rationnelle, on a nécessairement $2 < M^2$; la fonction carrée étant régulière, c'est donc qu'il existe un écart e irréductible entre M et les éléments de D.
(b) : Prenons alors un dénominateur d tel que $1/d < e$ – ici, $d = 6$. En avançant pas à pas, on trouvera nécessairement un numérateur n – et donc un rationnel $a = n/d$ – tel que a majore D tout en étant strictement inférieur à M. Ce dernier n'est donc pas « le plus petit des majorants ».
(c) : Dans $\mathbb{Q}$, il est impossible de mettre un "bouchon hermétique" « $\sup D$ » sur l'ensemble D. L'ensemble des réels palliera à cette lacune en vérifiant l'*axiome de la borne supérieure.*

Construction formelle Inutile de préciser que cette "définition" auto-référente n'en est pas une... Pour exprimer les réels à partir de formules sur $\mathbb{Q}$, on recourra à la construction suivante :

$$\begin{aligned}
\mathbb{R} = \big\{ A \subseteq \mathbb{Q} \mid \forall\, x \in A, \forall\, y \in \mathbb{Q},\ y < x \Rightarrow y \in A \qquad & \text{« } A \text{ est un intervalle infini à gauche »} \\
\wedge\ \sup A \text{ existe dans } \mathbb{Q} \ \Rightarrow \neg\,(\sup A \in A) \qquad & \text{« } A \text{ est ouvert à droite »} \\
\wedge\ \neg\,(A = \emptyset) \qquad & \text{« } A \text{ est non vide, i.e. } -\infty \notin \mathbb{R} \text{ »} \\
\wedge\ \neg\,(A = \mathbb{Q}) \qquad \big\} \qquad & \text{« } A \text{ est non plein, i.e. } +\infty \notin \mathbb{R} \text{ »}
\end{aligned} \tag{8.83}$$

Un réel est ainsi la donnée d'une manière de diviser $\mathbb{Q}$ en deux intervalles disjoints – une partie A et son complémentaire A^c : $\mathbb{R}$ sera l'ensemble des *coupures* de $\mathbb{Q}$, et chaque réel ensembliste A pourra être identifié avec le "point de découpe" $\sup A$. Tandis que l'ensemble D de l'équation (8.81) sera par définition le modèle ensembliste du réel $\sqrt{2}$, les rationnels seront naturellement représentés via l'injection canonique

$$\widetilde{\cdot} \ :\ \mathbb{Q} \to \mathbb{R} \tag{8.84}$$

$$q \mapsto \widetilde{q} = \{\, x \in \mathbb{Q} \mid\ x < q \,\}. \tag{8.85}$$

L'ordre usuel « $<$ » est simplement donné par l'inclusion, qui prolonge l'ordre usuel sur $\mathbb{Q}$ à $\mathbb{R}$ tout entier :

$$\forall\ q_1, q_2 \in \mathbb{Q},\ \big(q_1 < q_2 \iff \widetilde{q_1} < \widetilde{q_2} \quad \text{i.e.} \quad \widetilde{q_1} \subseteq \widetilde{q_2} \big). \tag{8.86}$$

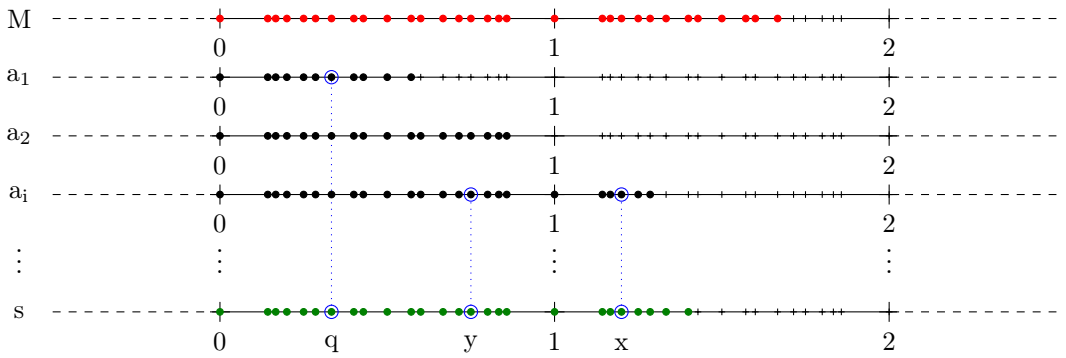


FIGURE 8.4 – La construction des réels présentée équation (8.83) vérifie bien l'axiome de la borne supérieure. On représente ici une famille de réels $(a_i)_{i\in I}$ majorée par le réel M. L'union ensembliste s des a_i est elle-aussi un réel, et est le plus petit majorant des a_i pour l'ordre donné par l'inclusion. Les éléments q, y et x sont ici placés pour illustrer la preuve en bas de page.

Théorème de la borne supérieure La définition formelle de l'addition et du produit sur les coupures de $\mathbb{Q}$ est malheureusement trop fastidieuse pour être développée ici... Nous terminons tout de même dignement notre chapitre de constructions ensemblistes par une preuve de la *complétude* de la structure définie équation (8.83) : $(\mathbb{R}, \subseteq)$ vérifie l'axiome (Sup) de l'équation 8.80.

Pour le montrer, on réécrit pas à pas l'énoncé de la-dite propriété.
Prenons donc $A = \{\, a_i, \text{ pour } i \in I \,\}$ une famille de réels que l'on suppose :

non vide i.e. I est non vide – il y a au moins un réel a_1 dans A.

majorée i.e. il existe un réel M tel que $\forall\, i \in I,\, a_i \subseteq M$.

Il s'agit de trouver un réel s tel que :

s **majore** A i.e. $\forall\, i \in I,\, a_i \subseteq s$.

s **soit plus petit que les autres majorants** i.e. si M est un majorant de A, alors $s \subseteq M$.

Proposer un candidat ensembliste n'est pas difficile : il suffit de considérer

$$s = \bigcup_{i\in I} a_i = \{\, q \in \mathbb{Q} \mid \exists\, i \in I, q \in a_i \,\} \tag{8.87}$$

qui est bien le plus petit des majorants de A pour l'inclusion – exercice ! La partie technique de la preuve est en fait de montrer que s est bien un *réel*, au sens de la formule (8.83) :

s **est un intervalle infini à gauche** Soit $x \in s$, $y \in \mathbb{Q}$ tels que $y < x$. Par définition de s, on dispose d'un indice i tel que $x \in a_i$: comme a_i est un réel, il est un intervalle infini à gauche, et donc $y \in a_i$, puis $y \in s$; Cqfd.

s **est ouvert à droite** Supposons par l'absurde que $\sup s$ existe dans $\mathbb{Q}$, et qu'il soit élément de s. Par définition de s, on dispose d'un nouvel indice j tel que $\sup s \in a_j$. Comme a_j est inclus dans s, $\sup s$ est aussi un majorant de a_j, et est donc le plus grand élément de ce dernier, son sup rationnel... Qu'il appartienne à a_j est alors contradictoire avec le fait que a_j soit ouvert à droite ; Cqfd.

s **est non vide** On a demandé l'existence d'un réel a_1 dans A : celui-ci étant non vide, on est donc assuré de l'existence d'un rationnel $q \in a_1$ i.e. $q \in s$; Cqfd.

s **est non plein** On a demandé l'existence d'un réel M qui majore A : on a alors $s \subseteq M \subsetneq \mathbb{Q}$; Cqfd.

s est donc un réel bien défini : on a démontré le "théorème" de la borne supérieure.

Conclusion

Au prix de bien des efforts, nous avons modélisé la paradoxale théorie des nombres réels dans une théorie des ensembles notoirement "rustique", mais fiable. Ce chapitre aura été l'occasion de nous frotter à de véritables preuves formelles et d'en apprécier la rigueur ; surtout, il nous aura permis d'apprécier l'effort conceptuel que l'on doit fournir pour définir la notion de « nombre », passant dans un premier temps des ensembles aux nombres entiers, puis des rationnels aux réels. Une approche mathématique rigoureuse aura donc permis de révéler les nombreux paradoxes qui se cachent derrière la notion a priori intuitive de *nombre réel*. C'est d'autant plus remarquable que le passage des nombres réels aux nombres *complexes* – dont on tend souvent à faire une montagne – se révélera en fait n'être guère plus problématique que son "analogue" additif, la transition des entiers naturels aux entiers relatifs.

Références

Pour écrire la section 8.1, je me suis basé sur les chapitres 2 et 3 du cours de *Logique et théorie des ensembles* de Patrick Dehornoy (donné à l'ENS en 2006). Vous pourrez le trouver à l'adresse suivante : `http://www.math.unicaen.fr/~dehornoy/surveys.html`. Un petit accroc tout de même : Plutôt que de travailler avec une logique du second ordre (quantification sur les formules), j'ai préféré exprimer l'axiome d'induction par une collection infinie dénombrable d'axiomes de récurrence ($\mathrm{Rec}\,\varphi$) du premier ordre portant sur les entiers – suivant en cela Martin Hils – afin de nous épargner une subtile discussion sur la *signature* d'un langage.

Le reste du chapitre relève quand à lui de la "culture générale" de classe préparatoire – les sections 3.2 et 3.3 donneraient matière à de très bons exercices de colle.

Chapitre 9

Histoire du calcul différentiel

Séance 12

À réécrire.

Dérivées et intégrales : après les complexes, le deuxième sommet des mathématiques de terminale. Nous adopterons un plan en six temps, six "grands noms" qui permettent de repérer commodément les avancées conceptuelles majeures :

Cavalieri, qui systématisa la découpe de figures géométriques en bandes "*indivisibles*".
Archimède, qui sema les ferments de la *généralisation* avec sa méthodes des *pesées*.
Newton, qui *systématisa* les raisonnements géométriques d'Archimède.
Leibniz, qui en fournit l'*abstraction* symbolique : le calcul différentiel.
Schwartz, qui *compléta* la théorie des fonctions dérivables.
Galerkine, qui permit de résoudre efficacement les problèmes numériques en *ingénierie*.

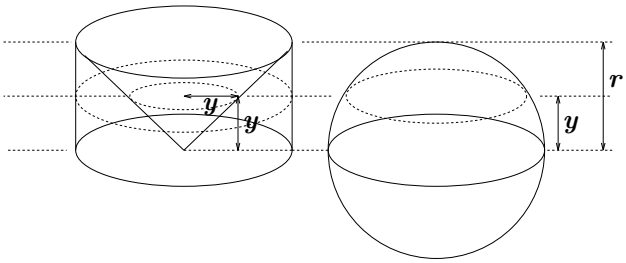
Inutile de le préciser : cette liste est à la fois anachronique et simpliste. La grande idée du calcul différentiel, sans doute la plus féconde de toute l'histoire des sciences, n'est pas l'œuvre d'un seul homme. Au delà des quelques héros immortalisés par nos manuels, elle résulte d'un échange permanent entre problèmes concrets et expériences de pensée, d'une lente percolation des idées ; d'une recherche vers la compréhension de l'infinitésimal. J'espère vous donner ici les clefs de lecture de cette belle histoire.

FIGURE 9.1 – Portraits de Cavalieri, Archimède (sur la médaille Fields), Newton, Leibniz, Schwartz et Galerkine. Images tirées de Wikipédia et du site `www.bibmath.net`.

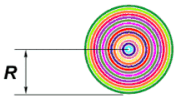
À réécrire.

Indivisibles de Cavalieri

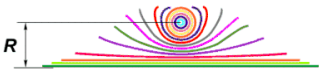
(a) Deux piles de pièces identiques ont même volume, quel que soit leur agencement.



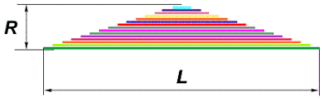
(b) Par une découpe en tranches astucieusement choisie, on peut ramener le calcul du volume d'une sphère à celui du cône : ici, chaque tranche de sphère a même aire que la tranche du "cylindre privé de cône" de même altitude.



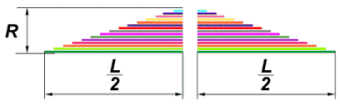
(c) Découpe d'un disque de rayon R en indivisibles concentriques, colorés ici.



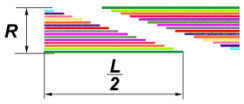
(d) Le périmètre d'un cercle croît linéairement avec son rayon : on déplie cette découpe.



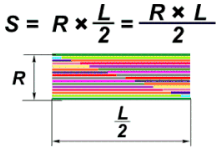
(e) Jusqu'à obtenir un triangle de hauteur R et de base $L = 2\pi R$, par définition de π.



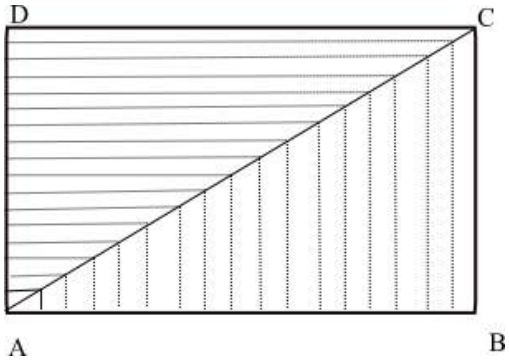
(f) On découpe ce triangle.



(g) Et on le réarrange...



(h) Pour connaître son aire : $S = \pi R^2$.



(i) Paradoxe des indivisibles. Le triangle ADC pouvant être décomposé en indivisibles plus *longs* que ceux de ABC, il devrait être plus étendu... ce qui est manifestement faux. Bien sûr, le problème est ici que l'*épaisseur* des petits éléments d'aire, ignorée par Cavalieri, n'est pas la même sur les deux découpes. Le calcul infinitésimal de Newton et Leibniz, dont les ferments avaient été déposées dès l'antiquité par Archimède, permettra de lever ce paradoxe par la formule du *changement de variables*.

FIGURE 9.2 – Illustration du principe de Cavalieri. Très populaire au XVII^{e} siècle, celui-ci permet d'identifier l'aire ou le volume de deux figures géométriques à condition de pouvoir les décomposer en *indivisibles* de mêmes longueurs. Toutes ces images sont tirées de Wikipédia : l'intuition (a) est due à Chiswick Chap, et le calcul du volume d'une sphère (b) à Michael Hardy ; l'animation du calcul de l'aire d'un disque (c-h) est quand à elle l'œuvre de Kaidor.

Pesée d'Archimède

« Qui lit Archimède et Apollonius en admirera moins les réussites de leurs successeurs. » Ma foi, je ne saurais pas vous dire si cette citation de Leibniz est apocryphe ou non... Mais ce qui est certain, c'elle qu'elle contient beaucoup de vrai ! Depuis la découverte du palimpseste d'Archimède en 1906 et de son traité inédit jusqu'alors, *De la méthode*, on ne peut qu'être soufflé par tant de génie : Quelle modernité ! C'est qu'en sus de la classique méthode d'exhaustion, le grand savant sicilien avait entrevu le dépassement des indivisibles "de Cavalieri" par une méthode pseudo-mécaniste, dite des *pesées*. Découvrir son exposition ne peut que nous laisser songeurs : si la deuxième guerre punique n'avait pas frappé Syracuse de plein fouet, le calcul différentiel aurait-il été découvert avec deux mille ans d'avance ?

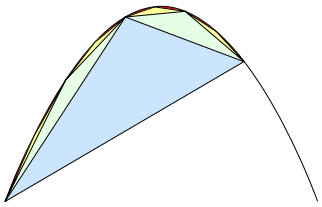
Esquisse de preuve Le raisonnement d'Archimède, illustré Figure 9.3b, est le suivant. Prenons un arc de parabole d'axe vertical, délimité par sa corde AB, et nommons C le point d'intersection entre la tangente à la parabole en A, et la verticale passant par B. Il s'agit de démontrer que l'aire $\mathbf{S}$ de l'arc de parabole est égale au tiers de celle de ABC, notée $\mathbf{T}$.

Pour cela, on imagine un système de levier, avec balancier AOA' : ABC est suspendu par un système de fils sans masses au fléau AO, tandis qu'une baguette de surface $\mathbf{S}'$ est placée en $\mathbf{A}'$.
Premier point clé, résultat classique de géométrie euclidienne : la masse de ABC répartie le long du fléau AO est équivalente à la masse (ou surface) totale concentrée en son centre de gravité – le *barycentre* – situé à la verticale de H_g, qui est tel que $OA = 3 \cdot OH_g$. Par le principe des leviers à l'équilibre, on a donc $\mathbf{S}' = \frac{1}{3}\mathbf{T}$.
Cœur de la méthode : on peut, de manière équivalente, imaginer que le triangle ABC est la juxtaposition de k baguettes trapézoïdales comme $MNPQ$, suspendues à la verticale de leurs centres de gravités H. À chacune de ces tranche doit donc correspondre un "bout de baguette $\mathbf{S}'$" : ici, $mm'n'n$. On peut alors écrire $\mathbf{S}'$ comme la somme des aires de ces bouts $mm'n'n$.
Point technique : par un argument de géométrie euclidienne ad hoc – Archimède n'a pas d'expression analytique de l'arc de parabole, mais une simple définition géométrique –, on montre que l'aire de $mm'n'n$ est comprise entre celle des deux trapèzes encadrant la tranche de parabole, $MRrs$ et $MNRS$. On a donc

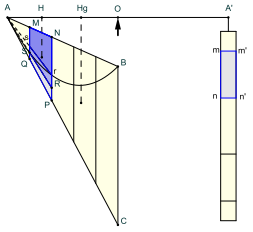
$$\sum \text{Aire}(MNrs) \leqslant \sum \text{Aire}(mm'n'n) = \mathbf{S}' = \frac{1}{3}\mathbf{T} \leqslant \sum \text{Aire}(MNRS), \tag{9.1}$$

$$\text{et} \quad \sum \text{Aire}(MNrs) \leqslant \mathbf{S} \leqslant \sum \text{Aire}(MNRS). \tag{9.2}$$

En augmentant le nombre de tranches verticales k, $\sum \text{Aire}(MNrs)$ et $\sum \text{Aire}(MNRS)$ pourront être rapprochées arbitrairement près : c'est donc nécessairement que $\mathbf{S} = \mathbf{S}' = \frac{1}{3}\mathbf{T}$.



(a) Preuve par exhaustion : un raisonnement classique attribué à Eudoxe de Cnide, où l'aire d'une figure est encadrée par celle de deux séries de triangles : l'une légèrement en excès, l'autre légèrement en retrait. En faisant tendre le nombre de triangles vers l'infini, les deux aires convergent, et on obtient par pincement l'aire de la figure "lisse". C'est l'idée qui sera au cœur de la construction de l'intégrale par Riemann.



(b) Preuve par pesée : très innovant, ce raisonnement permet d'aller "au-delà" des indivisibles sans épaisseur. Son statut épistémologique est l'objet d'une grande controverse : Archimède tenait-il cette preuve "mécaniste" pour rigoureuse ? S'agissait-il à ses yeux d'un simple guide pour l'intuition, ou bien d'une méthode révolutionnaire découverte au crépuscule de sa vie ?

FIGURE 9.3 – Deux démonstrations, par Archimède, de la quadrature de la parabole, i.e. du calcul de l'aire d'un bout de parabole délimité par sa corde. Ces deux figure sont tirées de Wikipédia : (a) est due à Jim Belk, Pbroks13, et (b) à HB. On pourra lire à ce sujet *Œuvres d'Archimède, traduites littéralement, avec un commentaire, par F. Peyard (1807).*

À réécrire.

Le calcul différentiel

Malheureusement, Archimède mourut lors du sac de Syracuse en 212 avant notre ère, et ses intuitions ne furent pas poussées plus avant. Pour passer d'un raisonnement novateur à une véritable méthode universelle, il fallut attendre les grands travaux des mathématiciens de l'âge classique, Descartes, Fermat, Newton et Leibniz en particulier.

Le calcul différentiel est une prise de conscience majeure. Celle d'une abstraction, le calcul infinitésimal, et de ses dérivées "$f' = \frac{\mathrm{d}f}{\mathrm{d}t}$" qui permettent de court-circuiter tous les raisonnements mécanistes, toutes les fulgurances géométriques par une simple liste de règles algébriques. Retracer le déroulement précis de cette révolution n'est pas une mince affaire : du vivant même de Newton et Leibniz, la paternité de cette invention devait mener à une amère controverse.

Au lecteur intéressé, je conseille l'excellent petit livre de Vladimir Arnold, *Huygens and Barrow, Newton and Hooke*. Nous nous contenterons ici d'une présentation archétypale des principaux événements : opposition entre Newton le géomètre et Leibniz le symboliste, avant la justification moderne de Weierstrass par la notion de limite rigoureusement définie.

De l'attraction gravitationnelle aux lois de Kepler

La principale contribution de Newton a été de comprendre que l'on pouvait *systématiser* les raisonnements géométriques des anciens. Dans ses *Principia Mathematica*, Newton démontre ainsi son résultat le plus fameux :

Théorème 9.1 (Dérivation par Newton des lois de Kepler)**.** *Supposons acquis les trois principes suivants :*

1. *« Tout corps persévère dans l'état de repos ou de mouvement uniforme en ligne droite dans lequel il se trouve, à moins que quelque force n'agisse sur lui, et ne le contraigne à changer d'état. »*
2. *« Les changements qui arrivent dans le mouvement sont proportionnels à la force motrice ; et se font dans la ligne droite dans laquelle cette force a été imprimée. »*
3. *« L'action est toujours égale à la réaction ; c'est-à-dire que les actions de deux corps l'un sur l'autre sont toujours égales et de sens contraires. »*

*Supposons de plus que dans son mouvement dans l'espace, la Terre (ou tout autre astre) soit uniquement soumise à une force d'*attraction gravitationnelle*, dirigée de la Terre vers le Soleil, de norme :*

$$F_{S\leftrightarrow T} = \frac{Gm_S m_T}{ST^2}, \tag{9.3}$$

où m_S et m_T sont les deux masses, et G un coefficient de proportionnalité universel.

Alors le mouvement de la Terre obéit aux trois lois de Kepler :

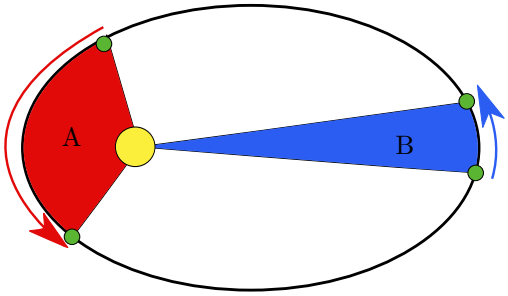
1. *Sa trajectoire est une ellipse dont le Soleil occupe l'un des foyers.*
2. *Dans deux intervalles de temps de même durée, le segment reliant la Terre au Soleil couvre des aires de mêmes surfaces.*
3. *Le carré de la période de révolution de l'astre (ou* année*) est lié par une loi de proportionnalité au cube du demi-grand axe de la trajectoire elliptique.*

Démonstration. Elle se décompose en trois temps.

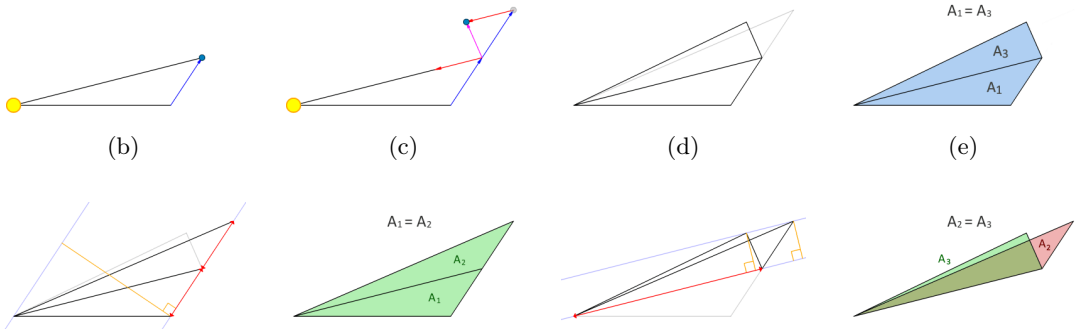
D'abord, comme indiqué Figure 9.4, on commence par montrer qu'une force *radiale* exercée du Soleil sur la Terre entraîne nécessairement la deuxième loi de Kepler.
Ensuite, on peut montrer qu'une loi en $1/ST^2$ implique que la trajectoire de l'astre est une *conique* dont le Soleil est un des foyers : ellipses pour les planètes, astéroïdes et comètes (trajectoire periodique), paraboles (cas limite) ou arcs d'hyperboles pour les corps étrangers au système solaire. Cette deuxième étape aura été rendue possible par les travaux de caractérisation des coniques au III$^{\text{e}}$ siècle avant notre ère par le rival d'Archimède, Apollonius de Perge.
Enfin, la proportionnalité de $F_{S\leftrightarrow T}$ vis-à-vis de m_T entraînera la troisième loi de Kepler. □

Le théorème ci-dessus est la preuve éclatante du génie mathématique de Newton. Il est d'ailleurs amusant de voir que son nom est resté attaché aux trois lois mécaniques dites "de Newton", qui sont en fait dues à ses prédécesseurs Galilée, Torricelli, Descartes, Huygens ou Hooke : elles ne figuraient dans son traité que comme de simples *rappels*. Le cœur de son œuvre est bien la dérivation ci-contre, la mise en évidence du lien entre une loi "en $1/r^2$" et des trajectoires elliptiques ; Newton ne s'en cachera pas, et écrira en 1676 à son maître Robert Hooke :

« If I have seen further it is by standing on the shoulders of Giants. »



(a) La deuxième loi de Kepler : dans deux intervalles de temps de même durée, le segment reliant un astre au Soleil couvre des aires de mêmes surfaces – ici, $A = B$. Une planète bouge donc d'autant plus vite qu'elle est proche du Soleil.



(b) (c) (d) (e) (f) (g) (h) (i)

FIGURE 9.4 – Preuve par Newton qu'une force centripète radiale entre le Soleil et la Terre induit la deuxième loi de Kepler, illustrée en (a). (b) Considérons le parcours d'une planète sur un intervalle de temps de longueur Δt : on représente ici les deux segments $ST_{t-\Delta t}$ et ST_t. (c) Supposons que sur l'intervalle de temps suivant, une force rouge *constante* soit appliquée sur T, selon la direction ST_t. Cela correspond à dévier la Terre d'une trajectoire "sans force" en translation rectiligne uniforme (d'après les travaux de Galilée), qui la mènerait à un point $\widetilde{T}_{t+\Delta t}$ (en gris), vers une position réelle $T_{t+\Delta t}$ (en bleu). (d) Trois triangles vont alors nous intéresser : $ST_{t-\Delta t}T_t$ d'aire A_1, $ST_tT_{t+\Delta t}$ d'aire A_3, et $ST_t\widetilde{T}_{t+\Delta t}$ d'aire A_2 (en gris). (e) Il s'agira de montrer que $\mathbf{A_1 = A_3}$. (f) On remarque que $ST_{t-\Delta t}T_t$ et $ST_t\widetilde{T}_{t+\Delta t}$ ont même base (en rouge) et même hauteur (en jaune). (g) Ils ont donc même aire. (h) De même, $ST_t\widetilde{T}_{t+\Delta t}$ et $ST_tT_{t+\Delta t}$ ont même base ST_t (en rouge), et même hauteur (en jaune), puisque la force appliquée entre t et $t+\Delta t$ sur T l'aura déviée en direction de ST_t. (g) On a donc $A_1 = A_2 = A_3$. En réduisant à l'infinitésimal l'intervalle de temps Δt, on trouve donc que si la force appliquée sur le point mobile T est à chaque instant dirigée selon l'axe ST_t, alors les petits triangles parcourus $ST_tT_{t+\Delta t}$ restent d'aire constante proportionnelle à Δt. En sommant ceux-ci, on retrouve bien la deuxième loi de Kepler. Images tirées de Wikipédia : (a) par Chatsam, et (b-i) par Lucas V. Barbosa.

À réécrire.

Abstraction par le calcul infinitésimal de Leibniz

Au delà de la simple généralisation des raisonnements géométriques "sur l'infiniment petit", l'innovation du *calcul infinitésimal* est l'abstraction de ceux-ci sous la forme d'un langage formel *algébrique*. Avant la justification rigoureuse, *axiomatique* de ces raisonnements – qui n'arrivera que dans les années 1860 avec les travaux de Karl Weierstrass –, deux modes de pensée vont s'opposer.

Notations de Newton Pour Newton, qui étudie l'évolution *par rapport au temps* de quantités notées x, y (coordonnées) ou encore S (surface), deux opérations sont primordiales : la "dérivation par rapport au temps", symbolisée par le point, et l'"intégration par rapport au temps", symbolisée par une apostrophe. Si x désigne la première coordonnée d'un point mobile, on convient par exemple que

— $\dot{x}$ désigne sa *vitesse*,

— $\acute{x}$ désigne ses sommes cumulées par rapport au temps. Si x est un débit d'eau à travers un tube, $\acute{x}$ est donc le volume total écoulé depuis un instant de référence.

À l'aide de raisonnements géométriques, on pourra alors montrer que ces opérations sont inverses l'une de l'autre : $\dot{\acute{x}} = x$. Ces notations sont très pratiques lorsqu'il s'agit d'exprimer des lois cinématiques : en coordonnées, la deuxième "loi de Newton" s'écrira par exemple

$$m\ddot{x} = F(x). \tag{9.4}$$

Malheureusement, si la dérivation par rapport au temps peut s'appuyer sur une forte intuition, la partie "intégration" reste malaisée à manipuler.

Notations de Leibniz La solution sera apportée par Leibniz. Plutôt que de s'appuyer sur de brillants raisonnements géométriques et cinématiques, celui-ci va proposer une vision purement algébriste, *calculatoire* de la dérivation, en postulant la cohérence d'un système de calcul fondé sur les *infinitésimaux*. Plus précisément, pour chaque variable d'intérêt comme x, y, z (coordonnées), t (le temps) ou S (la surface balayée), Leibniz introduira une *variation infinitésimale* associée : celles-ci seront notées $\mathrm{d}x$, $\mathrm{d}y$, $\mathrm{d}z$, $\mathrm{d}t$ ou encore $\mathrm{d}S$. Règle importante de ce nouveau calcul : dans une addition, un *infinitésimal* d'ordre 1 (i.e. un multiple de $\mathrm{d}t$, ...) sera toujours négligeable devant un terme d'ordre 0 (un terme "classique"). En "multipliant par $\mathrm{d}t$", un terme d'ordre 2 en $\mathrm{d}t^2$ sera lui aussi négligeable devant un terme d'ordre 1, et ainsi de suite.

Le point clé sera alors d'admettre qu'au voisinage de toute valeur t_0, une fonction $f(t)$ dépendant d'une variable t pourra être développée linéairement – ou *dérivée* au voisinage de t_0 :

$$f(t_0 + \mathrm{d}t) \;=\; f(t_0) + \frac{\mathrm{d}f}{\mathrm{d}t}(t_0) \cdot \mathrm{d}t + o(\mathrm{d}t), \tag{9.5}$$

où $o(\mathrm{d}t)$ désigne un terme d'ordre 2 ou plus, et où $\frac{\mathrm{d}f}{\mathrm{d}t}(t_0)$ est une fonction de t_0, appelée *dérivée de f par rapport à t en t_0*, que l'on peut aussi dériver sans sourciller. Si $x(t)$ est un point mobile dépendant du temps, $\frac{\mathrm{d}x}{\mathrm{d}t}$ désignera donc sa vitesse et $\frac{\mathrm{d}^2x}{\mathrm{d}t^2}$ son accélération.

Que penser alors de la somme d'une quantité au cours du temps ? Reprenons l'exemple de la deuxième loi de Kepler : il s'agissait de comprendre l'évolution de la surface balayée par le segment Soleil-Terre entre deux instants t_0 et t_1. En notant $S(t)$ la surface balayée depuis l'instant t_0, Leibniz propose tout simplement de calculer la variation de surface par rapport au temps $\frac{\mathrm{d}S}{\mathrm{d}t}$, puis de sommer ces taux de variations entre t_0 et t_1, on n'oubliera pas de pondérer le tout par l'*épaisseur* des tranches temporelles, $\mathrm{d}t$:

$$S(t_1) - S(t_0) \;=\; \int_{t=t_0}^{t_1} \frac{\mathrm{d}S}{\mathrm{d}t}(t)\ \mathrm{d}t \;=\; \int_{t=t_0}^{t_1} \mathrm{d}S(t), \tag{9.6}$$

où le signe « $\int$ » n'est rien d'autre qu'un « S » de « Somme » allongé, analogue continu du « $\sum$ » discret.

La dérivée du produit Pour bien mettre en valeur la différence d'esprit entre les deux écoles, intéressons-nous à la règle dite "du produit" et à sa version intégrée, l'*intégration par parties* qui traduit le taux d'évolution d'une quantité (xy) produit de deux variables x et y. Pour Newton, il s'agit d'une évidence géométrique : si

$$(\dot{xy}) = \dot{x}y + x\dot{y} \qquad \text{et donc} \qquad xy = (\dot{\acute{x}y}) + (\acute{x}\dot{y}) \text{ à une constante additive près,} \tag{9.7}$$

c'est à cause de la Figure 9.5 : l'infinitésimal "d'ordre 2" le long de la diagonale est manifestement négligeable devant celui des tranches. L'approche quasi-axiomatique de Leibniz permet quand à elle d'arriver à ce résultat sans le moindre dessin : il suffit d'écrire

$$\begin{aligned}
(xy)(t+\mathrm{d}t) &= x(t+\mathrm{d}t)\cdot y(t+\mathrm{d}t) && (9.8)\\
&= \left(x(t)+\frac{\mathrm{d}x}{\mathrm{d}t}\cdot\mathrm{d}t+o(\mathrm{d}t)\right)\cdot\left(y(t)+\frac{\mathrm{d}y}{\mathrm{d}t}\cdot\mathrm{d}t+o(\mathrm{d}t)\right) && (9.9)\\
&= x(t)y(t)+\frac{\mathrm{d}x}{\mathrm{d}t}\cdot\mathrm{d}t\cdot y(t)+x(t)\cdot\frac{\mathrm{d}y}{\mathrm{d}t}\cdot\mathrm{d}t+\frac{\mathrm{d}x}{\mathrm{d}t}\cdot\frac{\mathrm{d}y}{\mathrm{d}t}\cdot(\mathrm{d}t)^2+o(\mathrm{d}t) && (9.10)\\
&= (xy)(t)+\left(\frac{\mathrm{d}x}{\mathrm{d}t}\cdot y(t)+x(t)\cdot\frac{\mathrm{d}y}{\mathrm{d}t}\right)\cdot\mathrm{d}t+o(\mathrm{d}t), && (9.11)
\end{aligned}$$

car les multiples d'un $o(\mathrm{d}t)$ ou de $\mathrm{d}t^2$ sont négligeables devant $\mathrm{d}t$. On a donc bien

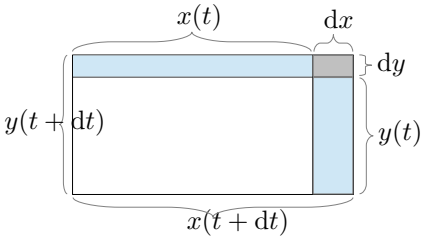
$$\frac{\mathrm{d}(xy)}{\mathrm{d}t} = \frac{\mathrm{d}x}{\mathrm{d}t}\cdot y + x\cdot\frac{\mathrm{d}y}{\mathrm{d}t}, \tag{9.12}$$

puis

$$(xy)(t_1)-(xy)(t_0) = \int_{t=t_0}^{t_1}\frac{\mathrm{d}(xy)}{\mathrm{d}t}\mathrm{d}t = \int_{t=t_0}^{t_1}\frac{\mathrm{d}x}{\mathrm{d}t}\cdot y\ \mathrm{d}t + \int_{t=t_0}^{t_1} x\cdot\frac{\mathrm{d}y}{\mathrm{d}t}\ \mathrm{d}t. \tag{9.13}$$

!!! Figure à dessiner !!!

(a) La sommation d'Abel, analogue discret avec tranches *finies* de la formule d'intégration par parties.



(b) À la limite, le terme d'angle $\mathrm{d}x\cdot\mathrm{d}y$ peut être négligé devant les termes de tranche $x\cdot\mathrm{d}y$ et $y\cdot\mathrm{d}x$. Image tirée de Wikipédia, par Nat Kuhn.

FIGURE 9.5 – La règle du produit peut être vue de manière purement géométrique. Si Newton se repose sur cette intuition pour appuyer ses raisonnements, Leibniz peut lui s'en détacher complètement : c'est ce qui assurera la primauté de son système de notations, qui triomphera sur le continent et finira même par s'imposer à Cambridge dans les années 1820.

À réécrire.

Premiers résultats La notation de Leibniz est une *bonne* notation, car elle permet de présenter comme des *évidences syntaxiques* des résultats qui sont en fait non-triviaux.

Proposition 9.1 ("Chain rule")**.** *Ainsi, la règle de dérivée d'une composée :*

$$\frac{\mathrm{d}(f\circ g)}{\mathrm{d}t}(t) = \frac{\mathrm{d}f}{\mathrm{d}g}(g(t))\cdot\frac{\mathrm{d}g}{\mathrm{d}t}(t), \tag{9.14}$$

que l'on pourrait presque interpréter comme une simplification de fraction.

Démonstration. On a :

$$\begin{aligned}
(f\circ g)(t+\mathrm{d}t) &= f(g(t+\mathrm{d}t)) && (9.15)\\
&= f\left(g(t) \quad + \frac{\mathrm{d}g}{\mathrm{d}t}(t)\cdot\mathrm{d}t + o(\mathrm{d}t)\right) && (9.16)\\
&= f(g(t)) \quad + \frac{\mathrm{d}f}{\mathrm{d}t}(g(t))\cdot\left(\frac{\mathrm{d}g}{\mathrm{d}t}(t)\cdot\mathrm{d}t + o(\mathrm{d}t)\right) && (9.17)\\
&= (f\circ g)(t) \quad + \left[\frac{\mathrm{d}f}{\mathrm{d}t}(g(t))\cdot\frac{\mathrm{d}g}{\mathrm{d}t}(t)\right]\cdot\mathrm{d}t + o(\mathrm{d}t). && (9.18)
\end{aligned}$$

D'où le résultat, par définition de la dérivée. □

Théorème 9.2 (Théorème fondamental de l'analyse)**.** *Soit $f(t)$ une quantité variable, et*

$$F : t_1 \mapsto \int_{t=t_0}^{t_1} f(t)\,\mathrm{d}t. \tag{9.19}$$

Alors

$$\frac{\mathrm{d}F}{\mathrm{d}t_1}(t_1) \;=\; f(t_1). \tag{9.20}$$

Démonstration. On a :

$$\begin{aligned}
F(t_1+\mathrm{d}t_1) &= \int_{t=t_0}^{t_1+\mathrm{d}t_1} f(t)\,\mathrm{d}t && (9.21)\\
&= \int_{t=t_0}^{t_1} f(t)\,\mathrm{d}t \;+\; \int_{t=t_1}^{t_1+\mathrm{d}t_1} f(t)\,\mathrm{d}t && (9.22)\\
&= F(t_1) \quad + \int_{t=t_1}^{t_1+\mathrm{d}t_1} f(t_1)\,\mathrm{d}t \;+\; \int_{t=t_1}^{t_1+\mathrm{d}t_1} (f(t)-f(t_1))\,\mathrm{d}t && (9.23)\\
&= F(t_1) \quad + \quad f(t_1)\cdot\mathrm{d}t_1 \quad + \int_{t=t_1}^{t_1+\mathrm{d}t_1} (0+o(1))\,\mathrm{d}t && (9.24)\\
&= F(t_1) \quad + \quad f(t_1)\cdot\mathrm{d}t_1 \quad + \; o(\mathrm{d}t_1). && (9.25)
\end{aligned}$$

D'où le résultat, en se reposant donc sur le fait que la variation de f soit d'ordre au moins 1 sur $[t_1, t_1+\mathrm{d}t_1]$. □

En posant simplement les définitions, on peut donc retrouver toutes les règles usuelles dont la dérivée de l'inverse ou celle des fonctions polynomiales – exercice !

$$\frac{\mathrm{d}(1/f(t))}{\mathrm{d}t} = -\frac{1}{f^2(t)}\frac{\mathrm{d}f}{\mathrm{d}t}(t) \qquad \text{et} \qquad \frac{\mathrm{d}(t^n)}{\mathrm{d}t} = n\cdot t^{n-1}. \tag{9.26}$$

Limites de Weierstrass et sommes de Riemann

Avec ses infinitésimaux, Leibniz a donné aux savants un formidable outil de calcul que compléteront les *séries entières* de Newton, Taylor et Euler. Mais comment rendre ses raisonnements rigoureux ? Après tout, "arbitrairement" proches de 0 sans jamais s'annuler, les infinitésimaux ne sont guère commodes à imaginer. La solution sera apportée au milieu du XIX^{e} siècle par deux mathématiciens allemands : Karl Weierstrass et Bernhard Riemann – en simplifiant beaucoup les choses !

Limites Au cœur de leurs idées, on trouvera la notion de *limite* qui permet d'exprimer les comportements infinitésimaux sans avoir à sortir du cadre numérique des réels. Formellement, on dira qu'une fonction $f(t)$ tend en t_0 vers une limite a, ce que l'on note

$$f(t) \xrightarrow{t \to t_0} a, \qquad \text{ou} \qquad \lim_{t \to t_0} f(t) = a, \tag{9.27}$$

si

$$\forall\, \varepsilon > 0, \exists\, \eta > 0, \forall\, t \in \mathbb{R}, \quad |t - t_0| < \eta \Longrightarrow |f(t) - a| < \epsilon. \tag{9.28}$$

Mon but n'est pas de vous présenter ici un cours d'analyse de niveau terminale/sup. Comme à mon habitude, je vous renvoie donc à l'excellent site de David Delaunay pour plus de détails : aujourd'hui, son cours intitulé *Limites et continuité* `http://mp.cpgedupuydelome.fr/cours.php?id=45861`. On se contentera ici de remarquer que pour I un intervalle de $\mathbb{R}$, et $f : I \to \mathbb{R}$ une fonction *lisse* à valeurs réelles sur ce domaine, on peut définir deux quantités de manière commode à partir de f.

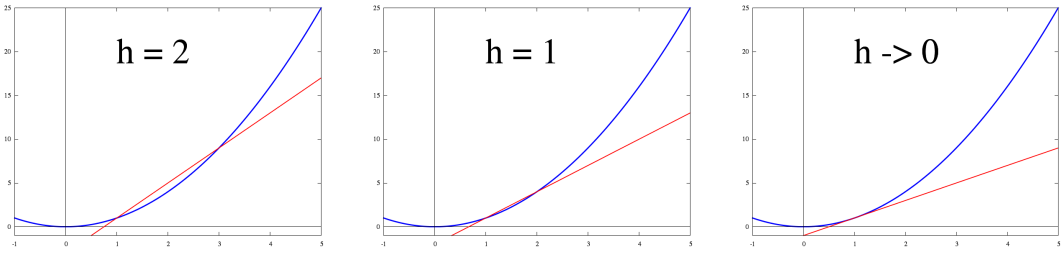
Différentiation D'abord, sa fonction dérivée ou *pente*, définie comme la limite infinitésimale des taux d'accroissements :

$$\forall\, x \in I,\ f'(x) \ = \ \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}. \tag{9.29}$$

Le coefficient $f'(x)$ caractérise la meilleure approximation *linéaire* de f au voisinage d'un point x de I : on a en effet, par définition de la limite,

$$\forall\, x \in I,\ f(x+h) = f(x) + f'(x) \cdot h + o(h), \tag{9.30}$$

où "$o(h)$" désigne une fonction de h qui tend vers 0 plus vite que tout multiple de h, i.e. telle que "$o(h)/h$" tende vers 0 avec h.



(a) $h = 2$ (b) $h = 1$ (c) $h \to 0$

FIGURE 9.6 – Définition de la dérivée comme limite du taux d'accroissement. Images tirées de Wikipédia, par IkamusumeFan.

À réécrire.

Intégration Ensuite, l'intégrale de f. Si $I = [a, b]$ est un intervalle borné de $\mathbb{R}$, on veut définir

$$\int_I f = \int_a^b f = \int_a^b f(t)\mathrm{d}t \tag{9.31}$$

comme l'aire sous la courbe de f sur I – éventuellement négative ou nulle, si f prend des valeurs négatives, ou si $a \geqslant b$. Le faire de manière rigoureuse pour toute fonction lisse – ou même seulement continue par morceaux – n'est pas évident : impossible en effet de toutes les étudier explicitement. Dans sa *construction de l'intégrale*, Bernhard Riemann procède en deux temps.

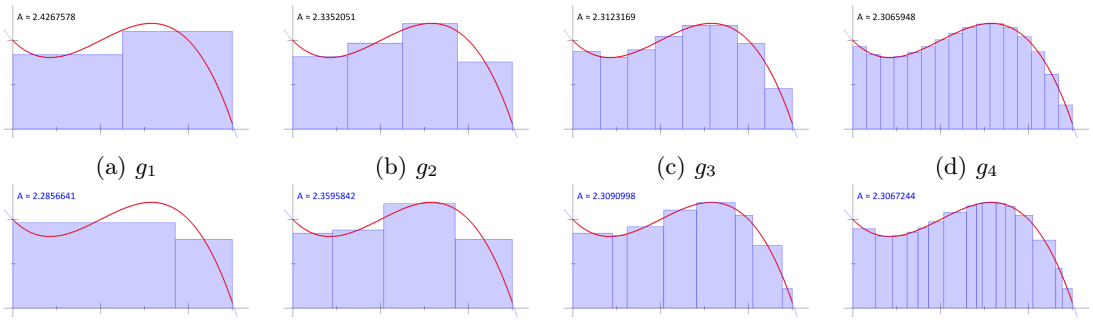
1. D'abord, il assigne une valeur à l'intégrale sur I de fonctions simples, les *fonctions en escalier* dont le graphe est une succession finie de rectangles : leurs aires seront faciles à calculer.
2. Ensuite, il montre que toute fonction suffisamment régulière peut être approchée *uniformément* par une suite de fonctions en escalier : si f est lisse sur I, alors il existe une suite de fonctions en escalier f_n telle que

$$\forall \varepsilon > 0, \exists\, n_0 \in \mathbb{N}, \forall\, n \geqslant n_0,\ \|f - f_n\|_{\infty, I} < \varepsilon \ \text{ i.e. } \ \forall\, x \in I, |f_n(x) - f(x)| < \varepsilon. \tag{9.32}$$

3. Enfin, il démontre que si f est limite uniforme de deux suites de fonctions en escalier g_n et h_n, alors les aires sous les courbes $\int_I g_n$ et $\int_I h_n$ convergent nécessairement, vers le même réel : ce sera par convention l'aire sous la courbe de la fonction limite f, le réel $\int_I f$ défini sans ambiguïté. Ce fait est illustré Figure 9.7.

Ces notations développées, on peut alors définir une fonction *intégrale* de f, associée à un point x_0 de I par la formule

$$I_{x_0}^f : x \mapsto \int_{x_0}^x f. \tag{9.33}$$



(a) g_1 (b) g_2 (c) g_3 (d) g_4

(e) h_1 (f) h_2 (g) h_3 (h) h_4

FIGURE 9.7 – Définition de l'intégrale de Riemann : toute fonction f (en rouge) régulière sur un intervalle peut être approchée uniformément par des suites de fonctions en escalier à subdivisions finies, qu'elles soient régulières (g_n, ligne du haut), ou irrégulières (h_n, ligne du bas). Fait remarquable : l'aire sous la courbe de ces suites converge nécessairement vers un réel qui ne dépend que de la fonction limite f : ce sera par convention l'*aire sous la courbe* de f. Images tirées de Wikipédia, par Kieff.

Premiers théorèmes Les définitions bien posées, on peut alors aborder les grands résultats d'analyse du lycée.

Théorème 9.3 (Théorème des accroissements finis)**.** *Soit f une fonction lisse sur I, et $a < b$ deux éléments de I. Alors*

$$\exists\, c \in\,]a,b[,\ \ f'(c) = \frac{f(b)-f(a)}{b-a}. \tag{9.34}$$

Autrement dit, tout accroissement *fini* de f entre deux points a et b trouve son analogue *infinitésimal* $f'(c)$ dans l'intérieur du domaine.

Démonstration. On considère la fonction lisse

$$g : x \mapsto f(x) - f(a) - \frac{f(b)-f(a)}{b-a} \cdot (x-a). \tag{9.35}$$

On a $g(a) = 0 = g(b)$: c'est donc que g est constante, ou bien que l'un de son minimum/maximum sur $[a,b]$ est atteint à l'intérieur du domaine. Dans les deux cas, on dispose d'un point $c \in\,]a,b[$ tel que

$$\forall\, x \in\ [a,b],\ g(x) \leqslant g(c) \qquad \text{ou} \qquad \forall\, x \in\ [a,b],\ g(x) \geqslant g(c). \tag{9.36}$$

Mais alors, que c soit l'antécédent d'un minimum ou d'un maximum de g, on aura $g'(c) = 0$. En effet, s'il s'agit par exemple d'un minimum, on aura que

$$\forall\, h < 0,\ \ \frac{g(c+h)-g(c)}{h} \leqslant 0 \qquad \text{donc à la limite, } g'(c) \leqslant 0, \tag{9.37}$$

$$\forall\, h > 0,\ \ \frac{g(c+h)-g(c)}{h} \geqslant 0 \qquad \text{donc à la limite, } g'(c) \geqslant 0, \tag{9.38}$$

et donc $g'(c) = f'(c) - \frac{f(b)-f(a)}{b-a} = 0$. □

Ce théorème – souvent passé sous le tapis dans les cours de terminale – est la clé de tout raisonnement qui passe des données infinitésimales à un résultat sur le comportement d'une fonction. Il permet ainsi de démontrer rigoureusement qu'une fonction dont la dérivée est positive sur un intervalle est nécessairement croissante – exercice !

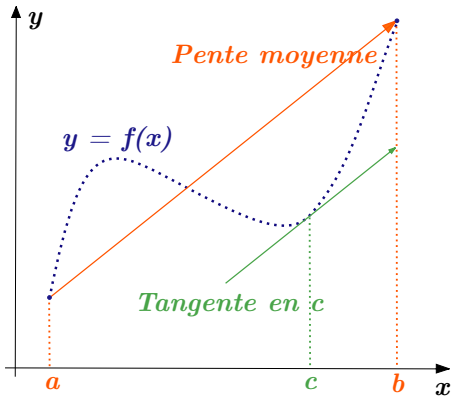


FIGURE 9.8 – Le théorème des accroissements finis : tout taux d'accroissement d'un point $(a, f(a))$ à un point $(b, f(b))$ peut être réalisé par un taux accroissement infinitésimal $f'(c)$, avec c compris entre a et b. Image tirée de Wikipédia, par DC2.

À réécrire.

Autre corollaire immédiat :

Théorème 9.4 (Unicité des primitives à une constante près)**.** *Soit F, G deux primitives de f sur I, i.e. deux fonctions telles que*

$$F' \;=\; f \;=\; G'. \tag{9.39}$$

Alors F et G sont égales à une constante près : $x \mapsto F(x) - G(x)$ est constante.

Démonstration. La dérivée est linéaire, par linéarité de la limite. Il s'agit donc de montrer que si $H = F - G$ est de dérivée partout nulle sur l'intervalle I, alors elle est constante. C'est bien vrai, en conséquence du théorème des accroissements finis : si a et b sont deux point de I, on sait qu'il existe c dans $]a, b[$ tel que

$$H(b) - H(a) = (b - a) \cdot H'(c) = 0, \quad \text{par hypothèse.} \tag{9.40}$$

Autrement dit, H est constante. □

Enfin, on peut démontrer rigoureusement le théorème fondamental de l'analyse :

Théorème 9.5 (Théorème fondamental de l'analyse)**.** *Soit $f : I \to \mathbb{R}$ ou $\mathbb{C}$ une fonction continue, x_0 un point de l'intervalle I, et*

$$I_{x_0}^f : x \mapsto \int_{x_0}^{x} f \tag{9.41}$$

une fonction intégrale de f.

Alors $I_{x_0}^f$ est dérivable, primitive de f :

$$\forall\, x \in I, \quad (I_{x_0}^f)'(x) = f(x). \tag{9.42}$$

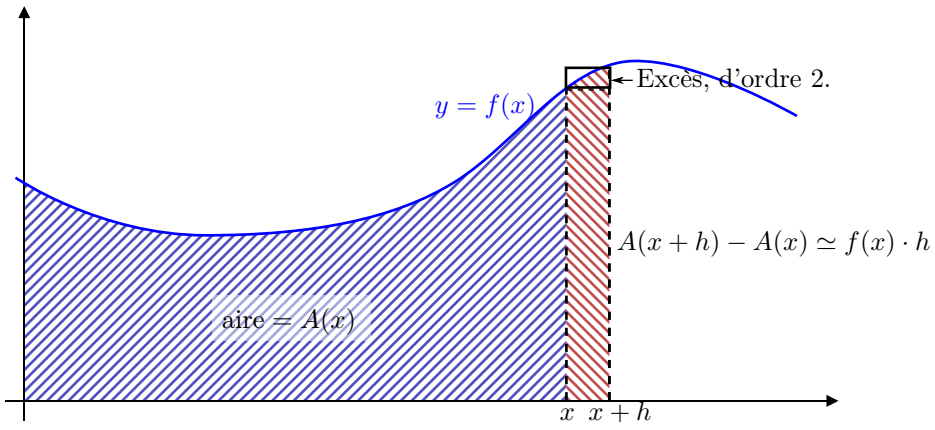


FIGURE 9.9 – Le théorème fondamental de l'analyse : le taux d'accroissements d'une fonction intégrale de f continue en un point t est exactement égal à $f(t)$. Une fonction *intégrale* est donc aussi une fonction *primitive*, ce qui l'identifie à une constante additive près par le théorème 9.4. Image tirée de Wikipédia, par Kabel.

Quadrature analytique de la parabole

Pour illustrer la puissance de ce nouveau calcul, reprenons la quadrature de la parabole illustrée Figure 9.3. Au début de ce chapitre, nous avions vu comment un génie de la trempe d'Archimède avait pu résoudre ce problème : à force d'efforts, une méthode d'*exhaustion* et une intuition pseudo-mécaniste – la méthode des *pesées* – avaient vu le jour, soutiens plus ou moins rigoureux de raisonnements géométriques complexes. Après deux mille ans de progrès conceptuels, ce calcul est maintenant à la portée de n'importe quel élève de terminale scientifique.

Paramétrisation du problème Suivant la méthode apprise en classe et introduite par Descartes, on commence par *paramétrer* le problème : d'une figure géométrique, on passe à un jeu d'équations. Comme illustré Figure 9.10b, on peut se doter d'un jeu de coordonnées orthonormé tel que la parabole soit la courbe d'équation

$$(P): \; y \;=\; p(x) \;=\; p\,x^2, \tag{9.43}$$

et que la corde d'intérêt soit celle passant par les points A et B de coordonnées respectives $(a, p\,a^2)$ et $(b, p\,b^2)$, où $a < b$ sont deux réels. Cette corde est donc la courbe d'équation

$$(C): \; y \;=\; c(x) \;=\; p\cdot\left(\frac{b^2-a^2}{b-a}(x-a)+a^2\right) \tag{9.44}$$

$$\;=\; p\cdot((b+a)\,x - ab)\,. \tag{9.45}$$

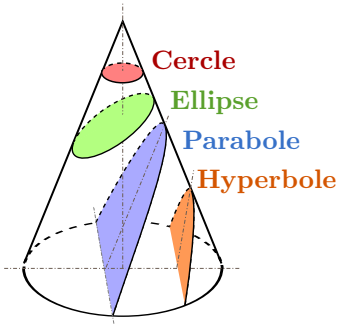
Enfin, la tangente à la parabole en A est une droite qui passe par $A = (a, p\,a^2)$, de coefficient directeur

$$\frac{\mathrm{d}p(x)}{\mathrm{d}x}(a) = \frac{\mathrm{d}(p\,x^2)}{\mathrm{d}x}(a) = 2\,p\,a. \tag{9.46}$$

C'est donc la droite d'équation

$$(T): \; y \;=\; t(x) \;=\; p\cdot\big(2\,a\,(x-a)+a^2\big) \tag{9.47}$$

$$\;=\; p\cdot\big(2\,a\,x - a^2\big)\,. \tag{9.48}$$



(a) Une parabole peut être définie géométriquement comme la section d'un cône par un plan suivant son angle d'inclinaison. Image tirée de Wikipédia, par Magister_Mathematicae.

!!! Figure à dessiner !!!

(b) La parabole est décrite par la fonction p, sa corde entre A et B par c, et sa tangente en A par t. Intégrer ces fonctions permettra de calculer directement l'aire S de l'arc de parabole et l'aire T du triangle délimité par t et c.

FIGURE 9.10 – Paramétrisation du problème de quadrature de la parabole.

À réécrire.

Calcul des aires Par le principe de Cavalieri, ou découpe des surfaces en tranches verticales, l'aire du grand triangle **T** est donnée par

$$\begin{aligned}\mathbf{T} &= \int_{x=a}^{b} [c(x) - t(x)]\,\mathrm{d}x && (9.49)\\ &= \int_{x=a}^{b} p\cdot\left[(b+a)\,x - ab - \left(2\,a\,x - a^2\right)\right]\,\mathrm{d}x. && (9.50)\end{aligned}$$

On a donc

$$\begin{aligned}\frac{1}{p}\,\mathbf{T} &= \int_{x=a}^{b} (b-a)\,x \quad + a\,(a-b)\,\mathrm{d}x && (9.51)\\ &= (b-a)\int_{x=a}^{b} x\,\mathrm{d}x \quad + a\,(a-b)\int_{x=a}^{b} 1\,\mathrm{d}x && (9.52)\\ &= (b-a)\left(\frac{b^2}{2} - \frac{a^2}{2}\right) + a\,(a-b)\,(b-a) && (9.53)\\ &= \frac{1}{2}\,(b-a)^3, && (9.54)\end{aligned}$$

car $x \mapsto \frac{1}{2}x^2$ est une primitive de $x \mapsto x$, elle-même primitive de $x \mapsto 1$.

De même, on trouve

$$\begin{aligned}\mathbf{S} &= \int_{x=a}^{b} [c(x) - p(x)]\,\mathrm{d}x && (9.55)\\ &= \int_{x=a}^{b} p\cdot\left[(b+a)\,x - ab - \left(p\,x^2\right)\right]\,\mathrm{d}x, && (9.56)\end{aligned}$$

puis

$$\begin{aligned}\frac{1}{p}\,\mathbf{S} &= \int_{x=a}^{b} -x^2 \;+\; (b+a)\,x \;-\; ab\;\,\mathrm{d}x && (9.57)\\ &= -\int_{x=a}^{b} x^2\,\mathrm{d}x \;+\; (b+a)\int_{x=a}^{b} x\,\mathrm{d}x \;-\; ab\int_{x=a}^{b} 1\,\mathrm{d}x && (9.58)\\ &= -\left(\frac{b^3}{3} - \frac{a^3}{3}\right) + \;(b+a)\left(\frac{b^2}{2} - \frac{a^2}{2}\right) \;-\; ab\,(b-a) && (9.59)\\ &= \frac{1}{6}\,(b-a)^3. && (9.60)\end{aligned}$$

On a donc retrouvé sans coup férir le résultat d'Archimède :

$$\mathbf{S} = \frac{1}{3}\,\mathbf{T} \qquad (9.61)$$

Conclusion Au travers de sa puissante abstraction symbolique, le calcul différentiel aura trivialisé la quasi-totalité des problèmes géométriques et physiques de l'antiquité. La porte est maintenant ouverte pour une science *moderne* de laquelle pleuvront les progrès techniques... Ainsi que de profonds problèmes posés par les mathématiques des grandes dimensions. C'est cette plongée en dimension infinie, l'analyse du XX$^{\text{e}}$ siècle, que je tâcherai de vous faire découvrir dans les deux chapitres suivants.

Chapitre 10

Dimension infinie, dualité et méthode des éléments finis

Séance 13

Formalisé au XVIII^{e} siècle, le calcul infinitésimal aura permis d'apporter une résolution méthodique à la quasi-totalité des problèmes mathématiques posés jusqu'alors. Pour cette science moderne débordant des frontières étriquées qui l'avaient jusqu'alors contenue, deux corridors allaient structurer les nouvelles avancées.

L'algèbre D'abors, les grands problèmes de *constructibilité* qui restaient toujours sans réponse. Hérités de la Grèce antique, ils posaient une question fort simple : est-il possible de construire telle ou telle figure géométrique à l'aide d'une règle et d'un compas ? Le doublement du cube : à partir d'un cube unité, un cube de volume double. La trisection de l'angle : à partir d'un angle arbitraire θ, un angle $\theta/3$. La quadrature du cercle : à partir d'un disque, un carré de même aire. Plus simple encore : on sait tous, depuis nos cours de rosaces, comment construire un hexagone régulier inscrit dans le cercle unité. Triangle, carré ne sont guère plus difficiles. À force d'astuce, on peut même construire un pentagone voire un polygone régulier à 17 côtés. Mais pourquoi, malgré tous nos efforts, n'arrivons-nous jamais au polygone à 9 côtés ?

Si le calcul différentiel permet de lier entre elles des longueurs et des aires, il n'apporte pas de réponse à ces questions qui portent, au fond, sur la structure *discrète* des équations polynomiales qui sous-tendent la géométrie classique. Un regard original, une théorie ad hoc devra donc être construite ; superbe, cette dernière pourrait à elle seule remplir un semestre du cours. Mais je ne m'étendrai pas plus à son sujet : trop éloignée de mes compétences de chercheur, elle sera bien mieux traitée par mes collègues algébristes.

L'analyse À l'opposé du spectre, on trouve les problèmes posés par la *dimension infinie*. C'est que pour résoudre des problèmes essentiellement paramétrés par une poignée de nombres réels (paraboles et ellipses ne sont jamais que la donnée de leurs grands axes), Leibniz et Newton ont ouvert la boîte de Pandore. Par le formalisme des infinitésimaux, ils ont étendu le calcul à *toutes* les courbes et fonctions lisses. Mais qu'est-ce, au juste, qu'une "fonction régulière" ? Peut-on parler rigoureusement d'objets que l'on ne peut pas décrire numériquement, de simples trajectoires "à main levée" ? Une plongée résolue dans le monde des fonctions "avec une infinité de paramètres" mettra vite le doigt sur des paradoxes inattendus. Il s'agira donc aujourd'hui de dépasser les insuffisances des *fonctions* naïves pour arriver à un formalisme véritablement adapté au calcul différentiel : celui des *distributions*.

Compacité en grande dimension

Avant tout, mettons au clair ce que l'on entend par "dimension" et "nombre de paramètres". Sans rentrer dans les détails formels, que l'on pourra trouver dans tout bon cours d'introduction aux espaces vectoriels (comme celui de David Delaunay, `mp.cpgedupuydelome.fr/cours.php?id=4240`), on se contentera d'identifier nos variables à des suites finies ou infinies de nombres réels. Un point x dans un espace de dimension d sera donc simplement assimilé à un d-uplet $(x_1, \dots, x_d) \in \mathbb{R}^d$ et on mettra l'accent sur la notion de *combinaison linéaire* : tout point x de l'espace peut s'écrire

$$x = \sum_{i=1}^{d} x_i \cdot e_i, \tag{10.1}$$

où $e_i = (0, \dots, 0, 1, 0, \dots, 0)$, avec un 1 en i^{e} position. On dira que la famille des e_i est une *base* de $\mathbb{R}^d$, car elle génère *tous* les points de l'espace, sans admettre d'équation d'annulation non triviale (impossible d'avoir $x = (0, \dots, 0)$ sans que tous les x_i soient nuls).

Le théorème de Bolzano-Weierstrass Un mathématicien voit l'ensemble $\mathbb{R}^d$ des d-uplets de nombres réels comme un espace de dimension d, où chacun des e_i est une "flèche" orthogonale aux autres. Partant de là, il est naturel de parler de *distance*, de *norme* dans $\mathbb{R}^d$. Si la valeur absolue reste la référence en dimension 1, plusieurs manières de la généraliser aux dimensions supérieures coexistent. On pourra, au choix, utiliser

$$\|x\|_1 = |x_1| + \cdots + |x_d|, \tag{10.2}$$

$$\|x\|_2 = \sqrt{x_1^2 + \cdots + x_d^2}, \tag{10.3}$$

$$\|x\|_\infty = \max\{|x_1|, \dots, |x_d|\}, \tag{10.4}$$

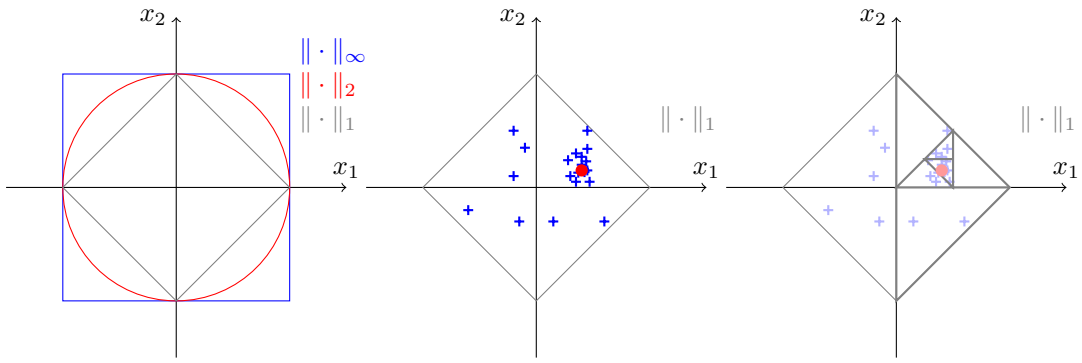
qui modélisent respectivement la distance "de Manhattan", la géométrie euclidienne et la convergence uniforme. Toutes ces normes sont *équivalentes* (en dimension finie), au sens où l'on peut toujours encadrer une boule pour la norme i par deux boules pour la norme j. Comme illustré Figure 10.1a, on pourra encadrer un "diamant" par deux cercles, ou un carré par deux diamants. Ainsi, une suite sera *bornée* ou *convergera* au sens d'une norme comme de l'autre – c'est un théorème classique de prépa.

Quelle que soit la dimension d ou la norme $\|\cdot\|_i$ choisie, un fait retiendra notre attention : la boule unité

$$B_i(0,1) = \left\{ x \in \mathbb{R}^d,\ \|x\|_i \leqslant 1 \right\} \tag{10.5}$$

est *compacte* ; au sens où toute suite $(u_n)_{n\in\mathbb{N}}$ y prenant ses valeurs admettra un *point d'accumulation*, une "limite partielle" u_∞ telle que pour tout rayon $\varepsilon > 0$, il existe une infinité d'indices n tels que $\|u_\infty - u_n\|_i \leqslant \varepsilon$. Illustrée Figure 10.1, cette propriété traduit une idée simple : dans une partie borné d'un espace de dimension finie, la "place" est comptée. Impossible, donc, d'y entasser une infinité de points régulièrement espacés.

Échantillonnage d'une boule en dimension finie : la malédiction de la dimension Mais que se passe-t-il lorsque la dimension de l'espace augmente ? Eh bien, de manière surprenante, la "place" à l'intérieur des boules fait de même. Chaque dimension accordant un degré de liberté supplémentaire, une augmentation de d permet de ranger toujours plus de points "régulièrement espacés" dans un rayon donné autour de l'origine. Connue sous le nom de "malédiction de la dimension" en théorie de l'apprentissage, une conséquence de ce fait est la grande difficulté

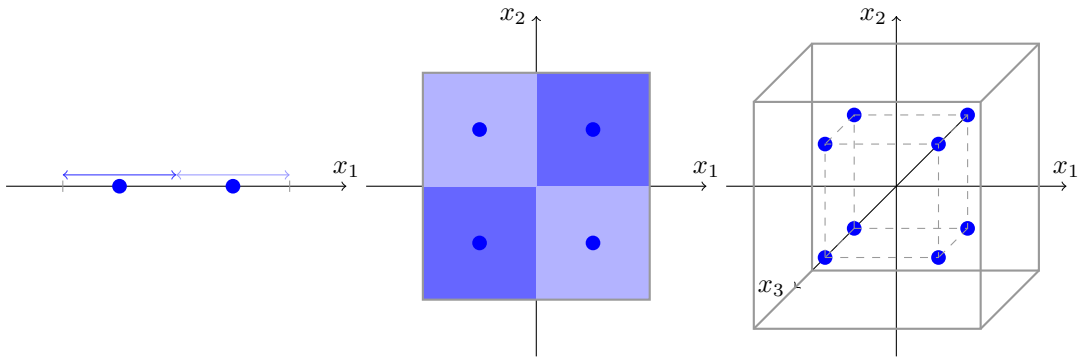


(a) Boules unités pour la "norme ∞", ($|x_1|$ et $|x_2| \leqslant 1$), la "norme 2", ($\sqrt{|x_1|^2 + |x_2|^2} \leqslant 1$) et la "norme 1", ($|x_1|+|x_2| \leqslant 1$).

(b) Une suite bornée (croix bleues) aura toujours une valeur d'adhérence...

(c) Son graphe présentera au moins un *point d'accumulation* (en rouge). Ici, une preuve par dichotomie.

FIGURE 10.1 – En dimension finie, les boules sont compactes. C'est le théorème de Bolzano-Weierstrass, dont on présente ici une preuve par dichotomie. De manière itérative, on construit une suite de triangles emboîtés, de diamètres tendant vers 0 et qui contiennent tous une infinité de termes de la suite – c'est le principe des tiroirs. L'intersection de tous ces triangles, nécessairement réduite à un unique point (par *complétude* de $\mathbb{R}$ et *séparation* de la norme), nous fournit alors la valeur d'adhérence de notre suite.



(a) Deux points à distance 1 l'un de l'autre dans $[-1, 1]$.

(b) Dans la boule de dimension 2, il y a plus de place.

(c) Et c'est encore plus vrai en dimension 3.

FIGURE 10.2 – En dimension d, on peut mettre jusqu'à $(1/r)^d$ points à distance r les uns des autres dans la boule de rayon 1. En ce sens, on peut donc dire que la boule de rayon 1 contient de l'ordre de 2^d boules de rayon $1/2$, 4^d boules de rayon $1/4$, etc. Par commodité, on utilise ici les boules pour la norme infinie ; mais le résultat reste vrai en normes 1 et 2 – à une constante multiplicative près.

pratique qu'il y a à "apprendre" une loi de décision dépendant de centaines de paramètres à partir d'une liste de quelques milliers/millions de décisions déjà connues (pensez à l'étiquetage d'une photo sur le web) : aussi grand que puisse paraître le nombre d'observations déjà étiquetées, il ne permet absolument pas d'échantillonner l'espace des paramètres avec une précision suffisante.

Le théorème de Riesz : en dimension infinie, des disques sans fond À la limite, en dimension infinie, il est même possible de construire des suites (u_n) telles que

$$\forall\, n \in \mathbb{N},\ \|u_n\| \leqslant 1 \qquad \text{et} \qquad \forall\, m, n \in \mathbb{N},\ \|u_m - u_n\| \geqslant r \tag{10.6}$$

pour toute valeur du paramètre r strictement inférieure à 1. Illustré Figure 10.3, ce théorème permet donc de produire des boules "sans fond", régions bornées de l'espace où l'on peut faire tenir une infinité de points bien espacés les uns des autres. C'est ce que j'appellerai « le paradoxe de Mary Poppins ».

Les fonctions forment un espace de dimension infinie

Notre petite étude à mis au jour un paradoxe étonnant... Mais est-il bien important ? Comment faire le lien avec les sujets qui nous préoccupent, l'analyse et la physique ?

Dimension d'un espace fonctionnel Il faut d'abord nous habituer à comprendre les fonctions comme des *vecteurs* dans un espace de *dimension infinie*. Pourquoi vecteur ? Parce qu'une fonction de $\mathbb{R}$ à valeurs dans $\mathbb{R}$ n'est jamais que la donnée d'une (infinité) de coordonnées $f(x)$. Pourquoi de dimension infinie ? Parce que ces coordonnées sont aussi nombreuses que les réels de la droite ; et que ces degrés de libertés sont bien souvent indépendants les uns des autres. Si on considère par exemple l'espace $E = C^\infty(\mathbb{R}, \mathbb{R})$ des fonctions lisses sur la droite réelle, et que l'on se donne $e_1, \ldots, e_n$ une famille de fonctions dans E, il n'est pas difficile de trouver une fonction f qui, croissant trop vite à l'infini, ne puisse être écrite comme une combinaison linéaire à coefficients réels des e_i. Impossible à munir d'une base de dimension n, pour tout entier n, E est donc un espace de dimension infinie.

Généralisation des normes aux espaces fonctionnels Comment parler de normes sur ces espaces dont les coordonnées ne peuvent pas toujours être énumérées ? Pour calculer $\|f\|$,

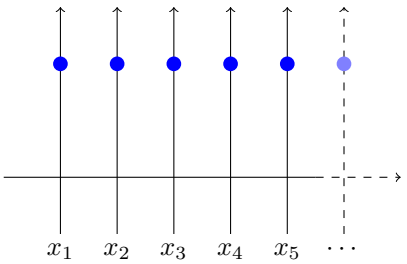


FIGURE 10.3 – Le lemme de Riesz : en dimension infinie, on peut trouver une suite infinie de points à distance 1 de l'origine, et tous éloignés les uns des autres par une distance arbitraire $r < 1$. Moralement, ces points peuvent être pensés comme étant chacun sur des axes différents, et sont les centres de boules de rayon $r/2$. La boule de rayon 2 contient donc une infinité de boules de rayon 0.45, ce qui est bien entendu impossible en dimension finie.

une simple sommation discrète des $|f(x)|$ ne peut convenir : à moins que f ne s'annule presque partout, on tombera forcément sur une somme infinie. Heureusement, le calcul intégral de Leibniz fournit une solution élégante à ce problème : il suffit de pondérer les coordonnées par les infinitésimaux « $\mathrm{d}x$ ».

Sur l'espace des fonctions continues de $[0,1]$ dans $\mathbb{R}$, on pourra ainsi définir les généralisations des normes vues aux équations (10.2-10.4) :

$$\|f\|_1 = \int_0^1 |f(x)|\,\mathrm{d}x, \tag{10.7}$$

$$\|f\|_2 = \sqrt{\int_0^1 |f(x)|^2\,\mathrm{d}x}, \tag{10.8}$$

$$\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|, \tag{10.9}$$

Une famille bornée sans valeur d'adhérence Mais alors, de nouveau illustré Figure 10.4, le « paradoxe de Mary Poppins » prend tout à coup une importance capitale. On y voit une suite de fonctions de densité, de norme 1 constante (charge totale Q conservée), sans valeur d'adhérence... Avec, pourtant, une "limite" intuitivement claire : la *distribution* de charges ponctuelle dite de Dirac, notée par les physiciens $Q \cdot \delta_0$, entièrement concentrée à l'extrémité droite du fil de fer.

Dans les espaces fonctionnels, le défaut de compacité des boules est devenu un enjeu formel de première importance. En obligeant le spécialiste à changer de description du monde (densités vs. masses ponctuelles) pour parler d'objets limites pourtant intuitifs, il entrave les calculs et empêche de systématiser les raisonnements. Dans la suite de ce chapitre, je vous propose donc de découvrir *le* bon cadre formel pour faire de la physique classique : la théorie des *distributions*, qui généralise celle des *fonctions* en garantissant une propriété de compacité (faible, non métrisable) des boules.

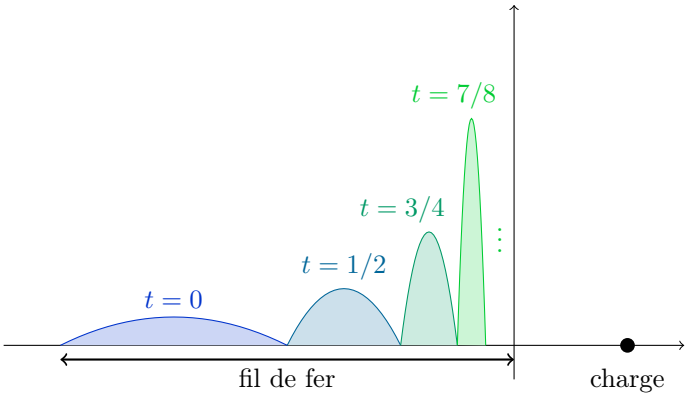


FIGURE 10.4 – Évolution fictive d'une densité de charges positives le long d'un fil de fer, attirées par une particule de charge négative sur la droite de l'image. À l'instant $t = 0$, on peut décrire la répartition de charges par une fonction de *densité*, en bleu : la charge totale le long d'une section $[a, b]$ du fil est égale à l'intégrale de la densité le long de celle-ci. Dans l'approximation naïve où les charges positives n'interagissent pas entre elles, on peut alors imaginer une évolution simpliste de la répartition de charges. Attirée, celle-ci se décalerait vers la droite en se contractant par effet de marée. En $t = 1$, toute la charge se trouverait concentrée à l'extrémité du fil... Mais alors, comment décrire cette évolution dans le formalisme des fonctions ? Une masse ponctuelle, parfaitement localisée, ne peut être représentée par une densité fonctionnelle.

Les distributions de Schwartz, fonctions généralisées

Pour échafauder notre théorie, on définit D l'espace des *fonctions tests* sur $\mathbb{R}$, ensemble des fonctions lisses (infiniment dérivables), à supports compacts. On dira de plus que deux fonctions tests sont *proches* si leurs supports et toutes leurs dérivées sont uniformément proches – l'énoncé formel est trop compliqué pour être énoncé ici. La remarque fondamentale de Laurent Schwartz est la suivante :

Lemme 10.1 (Plongement des fonctions dans D')**.** *Soit f une fonction continue sur $\mathbb{R}$. On peut lui associer une forme linéaire* continue *sur D,*

$$I_f : \varphi \in D \mapsto \int_{\mathbb{R}} f \cdot \varphi \in \mathbb{R}. \tag{10.10}$$

Alors f peut être identifiée à I_f, au sens où $I_f = I_g$ implique que $f = g$.

Démonstration. Il s'agit d'abord de remarquer que I_f est bien définie, linéaire, et continue.

Bonne définition Si φ est dans D, c'est qu'elle est lisse à support inclus dans un segment $[a, b]$, et donc bornée. $f \cdot \varphi$ est alors nulle en dehors de $[a, b]$, et continue : $I_f(\varphi) = \int_{\mathbb{R}} f \cdot \varphi = \int_a^b f \cdot \varphi$ est donc bien définie, réelle.

Linéarité Par linéarité de l'intégrale, on a

$$I_f(\lambda\varphi + \mu\psi) = \int_{\mathbb{R}} f \cdot (\lambda\varphi + \mu\psi) = \lambda \int_{\mathbb{R}} f \cdot \varphi + \mu \int_{\mathbb{R}} f \cdot \psi = \lambda I_f(\varphi) + \mu I_f(\psi). \tag{10.11}$$

Continuité Supposons disposer d'une suite de fonctions tests convergente au sens de D, $\varphi_n \to \varphi$. On a alors

$$|I_f(\varphi_n) - I_f(\varphi)| = |I_f(\varphi_n - \varphi)| = \left| \int_{\mathbb{R}} f \cdot (\varphi_n - \varphi) \right|. \tag{10.12}$$

Notons K le support de φ, et $\widetilde{K}$ un segment légèrement en excès. Alors, pour tout $\varepsilon > 0$, la définition de la convergence au sens des fonctions tests permet de trouver n_0 assez grand tel que pour tout n plus grand que n_0, on ait :

— $\varphi_n - \varphi$ nulle en dehors de $\widetilde{K}$ – convergence des supports.

— $|\varphi_n - \varphi|$ soit plus petit que ε sur tout $\widetilde{K}$ – convergence uniforme de la dérivée d'ordre 0.

On trouve alors que pour n plus grand que n_0,

$$|I_f(\varphi_n) - I_f(\varphi)| \leqslant \varepsilon \int_{\widetilde{K}} f. \tag{10.13}$$

Comme $\int_{\widetilde{K}} f$ est une quantité finie indépendante de ε, on a bien que $I_f(\varphi_n) \to I_f(\varphi)$, i.e. I_f séquentiellement continue sur D.

Reste à démontrer l'injectivité du procédé : si f et g sont deux fonctions continues telles que pour toute fonction test φ, $I_f(\varphi) = I_g(\varphi)$, il s'agit de montrer que $f = g$. Pour cela, supposons simplement qu'il existe x_0 dans $\mathbb{R}$ tel que $f(x_0) \neq g(x_0)$: quitte à intervertir f et g, supposons par exemple que $f(x_0) > g(x_0)$. Alors, par continuité de $f - g$, c'est qu'il existe un petit voisinage $]x_0 - \eta, x_0 + \eta[$ de x_0 où $f(x) - g(x)$ est supérieure ou égale à $(f(x_0) - g(x_0))/2$.

Le point clé est alors de remarquer que l'on peut construire une fonction test φ dans D qui soit positive et localisée sur $[x_0 - \eta, x_0 + \eta]$, une petite bosse qui est nulle partout, sauf au voisinage de x_0, où $f > g$. Mais alors :

$$\int_{\mathbb{R}} (f - g) \cdot \varphi \;=\; \int_{x_0 - \eta}^{x_0 + \eta} (f - g) \cdot \varphi \;>\; 0. \tag{10.14}$$

Autrement dit, $I_f(\varphi) > I_g(\varphi)$ en contradiction avec notre hypothèse de travail ; c'est donc qu'on a $f = g$. □

Le résultat précédent est fondamental : il permet d'identifier une *fonction continue* f à son action par intégration sur les fonctions test. À vrai dire, en analysant la preuve, on se rend compte que ce résultat peut être généralisé à l'ensemble des fonctions f *localement intégrables* – i.e. dont l'intégrale de la valeur absolue sur tout segment est finie, non divergente. Si f et g sont deux fonctions localement intégrables (ce qui garantit la continuité des formes linéaires I_f, I_g associées) telles que $I_f(\varphi) = I_g(\varphi)$ pour toute fonction test φ, alors f et g coïncident "*à un ensemble de mesure nulle près*" : f et g coïncident sur $\mathbb{R}$, sauf sur un ensemble "de longueur nulle", invisible pour l'intégrale. Cette petite distinction technique est là pour nous permettre d'identifier des fonctions "égales presque partout" – $\mathbb{1}_{[0,1]}$ et $\mathbb{1}_{]0,1[}$ par exemple – dont le comportement vis-à-vis de l'intégration est identique.

La morale à retenir est la suivante : **Toute fonction raisonnable, définie "à un ensemble négligeable près", peut être identifiée à son action sur les fonctions tests.** L'ensemble des fonctions "raisonnables" (localement intégrables) peut donc être vu comme une partie du *dual* de l'espace des fonctions tests, noté D'. Cet espace D', c'est celui des *distributions*. À l'avenir, on identifiera directement f à I_f, et on écrira

$$I_f(\varphi) = \int_{\mathbb{R}} f \cdot \varphi = \langle f, \varphi \rangle \,. \tag{10.15}$$

On étendra cette dernière notation à toute distribution : le *crochet de dualité* pourra donc être vu comme une généralisation de l'intégrale du produit (ou *produit scalaire*), avec à gauche une distribution et à droite une fonction test.

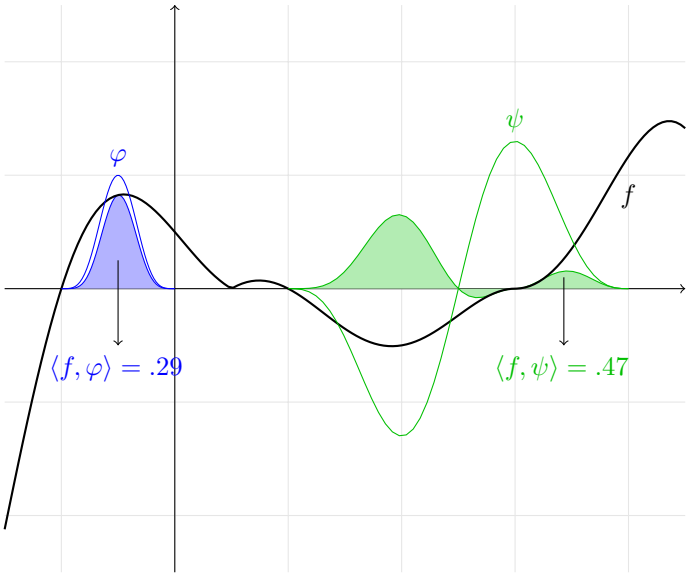


FIGURE 10.5 – Calcul du crochet de dualité entre une fonction f, en noir, et deux fonctions test φ et ψ. Il suffit de considérer les produits $f \cdot \varphi$ et $f \cdot \psi$, dont on retient les aires sous les courbes, indices réels de la corrélation entre la distribution f et le "test" lisse à support compact. Ici, on peut interpréter $\langle f, \varphi \rangle$ comme une moyenne de la masse de f autour de l'abscisse $-1/2$, et $\langle f, \psi \rangle$ comme un indicateur de la variation de f sur l'intervalle $[1, 4]$, ici positif car f y est globalement croissante. En utilisant des fonctions test toujours plus localisées, on pourra connaître f avec une précision arbitraire. Comme démontré au lemme 10.1, la donnée de tous les crochets $\langle f, \varphi \rangle$ caractérise donc la fonction f, à un ensemble de mesure nulle près.

Distributions qui ne sont pas identifiables à des fonctions A priori, cette identification entre f et I_f paraît dépourvue d'intérêt : pourquoi décrire une fonction par son *action* sur D, quand on dispose simplement de son graphe ? C'est que, comme démontré ci-dessous, l'ensemble des distributions est bien plus *gros* que celui des fonctions.

Proposition 10.1. *On définit le dirac en* 0, $\delta_0 \in D'$, *par son action sur les fonctions tests :*

$$\delta_0 : \varphi \in D \mapsto \varphi(0) \in \mathbb{R}. \tag{10.16}$$

Le dirac est bien linéaire et continu en φ : si φ et ψ sont proches, alors $\varphi(0)$ et $\psi(0)$ sont proches par définition de la proximité des fonctions tests. On a alors que :

— δ_0 *ne peut être représenté par une fonction f :*

$$\forall\ f \text{ fonction localement intégrable},\ \exists\ \varphi \in D,\ \langle \delta_0, \varphi \rangle \neq \langle f, \varphi \rangle . \tag{10.17}$$

— δ_0 *est la* limite simple *de toute suite de fonctions (f_n) définie de manière analogue aux bosses de la figure 10.4, par l'équation :*

$$\int_{\mathbb{R}} f_0 = 1 \qquad \text{et} \qquad f_n(x) = 2^n \cdot f_0(2^n \cdot x). \tag{10.18}$$

Formellement, on a :

$$\forall\ \varphi \in D,\ \ \langle f_n, \varphi \rangle \longrightarrow \varphi(0) = \langle \delta_0, \varphi \rangle . \tag{10.19}$$

Démonstration. Pour le premier point, supposons qu'il existe une fonction f telle que $\delta_0 = f$. On considère alors la suite de fonctions tests φ_n

$$\varphi_n : x \mapsto \varphi_0(2^n \cdot x), \tag{10.20}$$

où φ_0 est une "bosse" lisse centrée en 0, à support dans $[-1, 1]$, telle que $\varphi_0(0) = 1$. On a alors

$$\langle f, \varphi_n \rangle = \int_{\mathbb{R}} f \cdot \varphi_n = \int_{-1/2^n}^{+1/2^n} f(t) \cdot \varphi_n(t)\, \mathrm{d}t \ \xrightarrow{n \to +\infty}\ 0. \tag{10.21}$$

C'est bien sûr en contradiction avec le fait que $\langle \delta_0, \varphi_n \rangle = \varphi_n(0) = \varphi_0(2^n \cdot 0) = 1$.

Pour le second point, il s'agit d'un calcul simple (mais un peu technique, typique de la prépa) qui repose sur le fait qu'une fonction test φ est lisse, donc continue au voisinage de 0 : les moyennes locales $\langle f_n, \varphi \rangle$, de plus en plus resserrées, convergent nécessairement vers la valeur de φ en 0 . □

Travailler dans l'espace des *distributions* nous permet donc de parler de mesure ponctuelles, les *diracs*, dans un cadre formel englobant celui des fonctions : on a donné à la suite de fonctions f_n une limite. En fait, on peut montrer que le problème de *compacité des boules* a été résolu par le passage au dual, de l'espace des fonctions "gentilles" D à celui des distributions D' ; c'est-à-dire que toute suite "raisonnable" de fonctions admet une valeur d'adhérence. Il s'agit du théorème de *Banach-Alaoglu*, que je ne peux énoncer formellement ici faute de vocabulaire adéquat – topologie des espaces non métrisables, ou théorie des limites sans normes.

Interprétation économique Pour vous aider à comprendre la distinction entre distribution et fonction test, une petite analogie économique me semble appropriée. Si l'ensemble D des fonctions test était l'espace des *catalogues*, son dual D', espace des distributions, serait celui des *paniers* : à un catalogue φ, donnée pour tout produit x d'un prix $\varphi(x)$, un panier f associe bien un coût total $\langle f, \varphi \rangle$, réel, de manière linéaire en φ. Le lemme 10.1 peut alors se comprendre simplement : si deux paniers f et g nous mènent à payer le même prix pour toute répartition φ des tarifs, c'est que f et g étaient identiques.

De plus, si des distributions "à densité", plongement de fonctions continues, représentent des paniers "étalés", les diracs n'ont plus rien de mystérieux : ils correspondent simplement aux clients sûrs de leurs choix, intéressés par un seul produit.

Opérations usuelles On a plongé l'espace des fonctions – muni d'opérations simples – dans un espace plus gros, celui des distributions. Pour obtenir un cadre conceptuel satisfaisant, il est maintenant nécessaire de *prolonger* les opérations usuelles, des fonctions aux distributions. Il s'agit au fond du même travail que celui que nous avons accompli au chapitre 8 : là où le plongement de $\mathbb{Q}$ dans $\mathbb{R}$ avait permis de régler le problème de la *complétude*, le plongement de l'ensemble des fonctions dans celui des distributions nous aura permis de répondre aux paradoxes de la *compacité*. Eh bien, de même qu'il était agréable de savoir additionner, multiplier, comparer deux nombres réels, il va maintenant être indispensable de pouvoir additionner, redimensionner, dériver des distributions.

Addition Étant données deux distributions f et g, on peut simplement définir leur somme $f + g$ par la relation suivante :

$$\forall\ \varphi \in D,\ \langle f + g, \varphi \rangle\ =\ \langle f, \varphi \rangle + \langle g, \varphi \rangle\,. \tag{10.22}$$

Multiplication De même, si λ est un réel et f une distribution, le produit λf est immédiatement défini :

$$\forall\ \varphi \in D,\ \langle \lambda f, \varphi \rangle\ =\ \lambda\,\langle f, \varphi \rangle\,. \tag{10.23}$$

En poussant un peu plus loin, on peut définir le produit d'une distribution f et d'une fonction ψ infiniment dérivable :

$$\forall\ \varphi \in D,\ \langle \psi \cdot f, \varphi \rangle\ =\ \langle f, \psi \cdot \varphi \rangle\,. \tag{10.24}$$

C'est bien la seule définition compatible avec l'expression du crochet de dualité pour les distributions fonctionnelles : si f est une fonction, on a simplement

$$\forall\ \varphi \in D,\ \langle \psi \cdot f, \varphi \rangle\ =\ \int_{\mathbb{R}} \psi \cdot f \cdot \varphi\ =\ \int_{\mathbb{R}} f \cdot (\psi \cdot \varphi)\ =\ \langle f, \psi \cdot \varphi \rangle\,. \tag{10.25}$$

Maintenant, sera-t-il possible de définir le produit de deux distributions ? En toute généralité, **non**. Impossible en effet de définir le produit de deux diracs en 0 tout en restant fidèle à la notion de limite, la continuité du produit que l'on ne veut pas abandonner. En notant f_n une suite de fonctions lisses qui converge vers δ_0 définie à l'équation (10.18), on aurait en effet que pour φ une fonction test quelconque valant 1 en 0 :

$$\langle \delta_0 \cdot \delta_0, \varphi \rangle = \lim_{n \to +\infty} \langle \delta_0 \cdot f_n, \varphi \rangle \tag{10.26}$$

$$= \lim_{n \to +\infty} \langle \delta_0, f_n \cdot \varphi \rangle \tag{10.27}$$

$$= \lim_{n \to +\infty} (f_n \cdot \varphi)(0)\ =\ +\infty \qquad \text{si} \quad f_0(0) > 0. \tag{10.28}$$

Autrement dit, $\delta_0 \cdot \delta_0$ devrait assigner une valeur non réelle à tout fonction test ne s'annulant pas en 0 : c'est bien que $\delta_0 \cdot \delta_0$ ne peut être vue comme une *distribution*.

Comment dériver ce qui n'est même pas continu ?

En pratique, l'impossibilité de multiplier entre elles deux distributions quelconques n'est pas un vrai problème : les distributions n'ont pas été pensées pour construire des polynômes, mais pour nous permettre d'étudier des équations différentielles – ce sera le sujet de la fin du chapitre. L'important est donc de savoir *dériver* des distributions quelconques... Mais comment ? On a vu au chapitre précédent la formule d'intégration par parties :

$$\int_a^b f' \cdot \varphi \;=\; [f \cdot \varphi]_a^b - \int_a^b f \cdot \varphi'. \tag{10.29}$$

En passant à la limite sur a et b, on obtient donc la relation suivante :

$$\begin{aligned} \langle f', \varphi \rangle \;&=\; (f \cdot \varphi)(+\infty) - (f \cdot \varphi)(-\infty) - \langle f, \varphi' \rangle && (10.30)\\ &=\; -\langle f, \varphi' \rangle\,, && (10.31) \end{aligned}$$

car toute fonction test φ de D s'annule en dehors d'un segment. Voilà une formule qui caractérise uniquement la dérivée de f sans utiliser de limite de taux d'accroissements.

Dérivation Immédiatement généralisable, cette formule qui s'exprime au travers des seules fonctions test sera notre définition de la dérivée généralisée : pour f une distribution quelconque, on définit f' par

$$\forall\, \varphi \in D, \quad \langle f', \varphi \rangle \;=\; -\langle f, \varphi' \rangle\,. \tag{10.32}$$

L'astuce est brillante : pour définir une opération complexe comme la dérivée, il suffit de "faire porter le chapeau" aux fonctions tests – qui ont toutes les bonnes propriétés du monde. Mais ce qui n'est a priori qu'un simple jeu syntaxique sera en fait *la* bonne manière de manipuler nos objets, à la fois cohérente et intuitive.

Règle du produit On peut par exemple remarquer que la règle du produit tient toujours : si f est une distribution, ψ une fonction lisse et φ une fonction test quelconque, on a

$$\begin{aligned} \langle (\psi \cdot f)', \varphi \rangle &= -\langle \psi \cdot f, \varphi' \rangle && (10.33)\\ &= -\langle f, \psi \cdot \varphi' \rangle && (10.34)\\ &= -\langle f, (\psi \cdot \varphi)' \quad - \psi' \cdot \varphi \rangle && (10.35)\\ &= -\langle f, (\psi \cdot \varphi)' \rangle \;+ \langle f, \psi' \cdot \varphi \rangle && (10.36)\\ &= +\langle f', \psi \cdot \varphi \rangle \quad + \langle f, \psi' \cdot \varphi \rangle && (10.37)\\ &= +\langle \psi \cdot f', \varphi \rangle \quad + \langle \psi' \cdot f, \varphi \rangle && (10.38)\\ &= +\langle \psi \cdot f' \;+\; \psi' \cdot f, \varphi \rangle. && (10.39) \end{aligned}$$

Autrement dit,

$$(\psi \cdot f)' = \psi \cdot f' + \psi' \cdot f. \tag{10.40}$$

Dérivées de la valeur absolue Les égalités établies entre fonctions lisses sont toujours valables au sens des distributions : c'est exactement le sens du lemme 10.1. Plus intéressant, on peut maintenant établir de nouvelles identités jusqu'à présent rejetées comme "inaccessibles" – imaginaires ! Prenez par exemple la fonction valeur absolue :

$$\mathrm{abs} : x \in \mathbb{R} \mapsto |x| = \begin{cases} +x & \text{si } x \geqslant 0 \\ -x & \text{si } x \leqslant 0 \end{cases}. \tag{10.41}$$

On sait qu'elle n'est pas dérivable en 0 : la limite de ses taux d'accroissements à gauche y fait -1, tandis que ceux de droite y tendent vers $+1$. Sa courbe représentative n'admet donc pas de tangente à l'origine, mais présente plutôt un point anguleux. Heureusement, le formalisme des distributions nous permet maintenant de l'étudier comme une fonction à part entière, et d'en considérer les dérivées !

Dérivée d'un point anguleux Il suffit de procéder méthodiquement. Si φ est une fonction test, on aura

$$\begin{aligned}
\langle \mathrm{abs}', \varphi \rangle &= -\langle \mathrm{abs}, \varphi' \rangle && (10.42)\\
&= -\int_{\mathbb{R}} |t| \cdot \varphi'(t)\,\mathrm{d}t && (10.43)\\
&= -\left[\int_{-\infty}^{0} (-t) \cdot \varphi'(t)\,\mathrm{d}t \;+\; \int_{0}^{+\infty} (+t) \cdot \varphi'(t)\,\mathrm{d}t\right] && (10.44)\\
&= \int_{-\infty}^{0} t \cdot \varphi'(t)\,\mathrm{d}t \;-\; \int_{0}^{+\infty} t \cdot \varphi'(t)\,\mathrm{d}t. && (10.45)
\end{aligned}$$

Or, pour a et b deux bornes quelconques, on a par intégration par parties de la fonction lisse $t \mapsto t \cdot \varphi(t)$:

$$\int_a^b t \cdot \varphi'(t)\,\mathrm{d}t = [t \cdot \varphi(t)]_a^b - \int_a^b 1 \cdot \varphi(t)\,\mathrm{d}t. \tag{10.46}$$

En utilisant le fait que φ s'annule à l'infini – c'est une fonction test –, les termes de bords disparaissent et on trouve

$$\begin{aligned}
\langle \mathrm{abs}', \varphi \rangle &= -\int_{-\infty}^{0} 1 \cdot \varphi(t)\,\mathrm{d}t + \int_{0}^{+\infty} 1 \cdot \varphi(t)\,\mathrm{d}t && (10.47)\\
&= \int_{\mathbb{R}} \mathrm{sgn}(t) \cdot \varphi(t)\mathrm{d}t. && (10.48)
\end{aligned}$$

La dérivée de abs est donc identifiable à la fonction créneau sgn, définie par

$$\mathrm{sgn} : x \in \mathbb{R} \mapsto \mathrm{sgn}(x) = \begin{cases} +1 & \text{si } x > 0 \\ -1 & \text{si } x < 0 \end{cases}. \tag{10.49}$$

Rappelons que la valeur ponctuelle d'une fonction n'a aucune influence sur l'intégrale : la valeur assignée à sgn(0) n'a donc aucune importance.

Dérivée d'un créneau Pour calculer la dérivée du créneau sgn, la méthode est la même : pour φ une fonction test, on a

$$\begin{aligned}
\langle \mathrm{sgn}', \varphi \rangle &= -\langle \mathrm{sgn}, \varphi' \rangle && (10.50)\\
&= -\left[\int_{-\infty}^{0} (-1) \cdot \varphi'(t)\,\mathrm{d}t + \int_{0}^{+\infty} (+1) \cdot \varphi'(t)\,\mathrm{d}t\right] && (10.51)\\
&= \int_{-\infty}^{0} \varphi'(t)\,\mathrm{d}t - \int_{0}^{+\infty} \varphi'(t)\,\mathrm{d}t && (10.52)\\
&= (\varphi(0) - \varphi(-\infty)) - (\varphi(+\infty) - \varphi(0)) && (10.53)\\
&= 2\,\varphi(0), && (10.54)
\end{aligned}$$

en utilisant le théorème fondamental de l'analyse et le fait que φ s'annule en dehors d'un segment. Autrement dit,

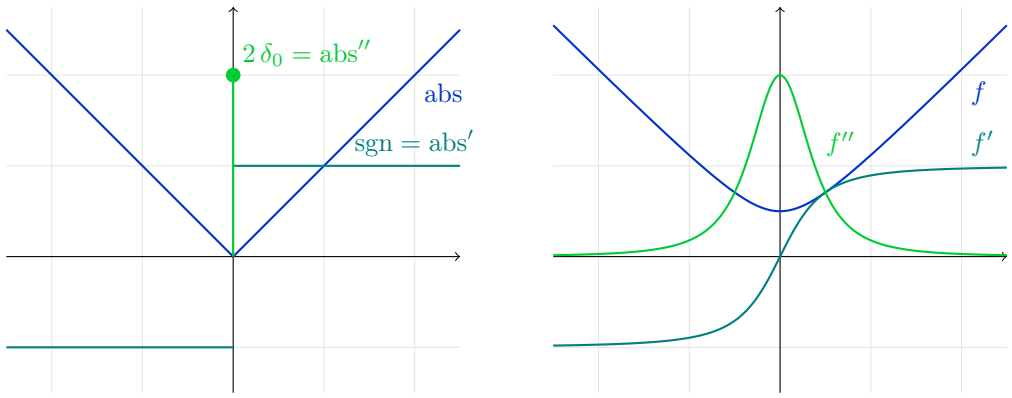
$$\mathrm{abs}'' \;=\; \mathrm{sgn}' \;=\; 2\,\delta_0. \tag{10.55}$$

Dérivée d'un dirac Contrairement aux apparences, rien ne nous force à nous arrêter là ! Les dérivées successives d'un dirac sont simplement définies par

$$\langle \delta_0', \varphi \rangle = -\varphi'(0), \qquad \langle \delta_0'', \varphi \rangle = +\varphi''(0), \quad \text{etc.} \tag{10.56}$$

Ce petit jeu symbolique nous aura permis de calculer à la fois *précisément* et *simplement* les dérivées de fonctions non-lisses présentant points anguleux ou créneaux. Fait remarquable, et essentiel pour l'utilité *pratique* de ces résultats, les calculs ci-dessus sont *stables* : si f_n est une suite de fonctions qui tend (au sens des distributions) vers abs, alors la suite des dérivées f_n' converge vers sgn, et la suite des dérivées secondes f_n'' vers $2\,\delta_0$. Si les fonctions f_n sont lisses, approchant au plus près le point anguleux de la valeur absolue en 0, les f_n' seront donc des "créneaux lisses" approchant sgn au plus près, tandis que les f_n'' seront de petites bosses convergeant vers un dirac en 0 de masse 2 – voir Figure 10.6.

Conclusion Le travail formel accompli par Schwartz aura permis de donner un sens rigoureux aux intuitions algébriques de Heaviside, Dirac et Sobolev. En pensant les fonctions non plus comme des formules ou des graphes, mais comme des *formes linéaires* sur un espace vectoriel de fonctions lisses, on aura réussir à faire "grossir" notre boîte à outils à moindres frais. Décrire une fonction par ses "empreintes digitales" facilement manipulables, les intégrales contre des fonctions tests, nous aura permis de trouver axiomatiquement la seule manière de "dériver un dirac". C'est ce point de vue proche de l'algèbre linéaire qui modèle aujourd'hui la recherche sur les équations différentielles.



(a) Dérivées successives de la valeur absolue, représentées de manière imagée.

(b) Dérivées successives d'une fonction f proche de la valeur absolue.

FIGURE 10.6 – La théorie des distributions permet de donner un sens approprié aux dérivées de fonctions irrégulières. Les formules obtenues sont souvent intuitives : elles avaient été utilisées sans justifications par des physiciens, une trentaine d'années avant la formalisation de Schwartz. En aplanissant ce terrain en friche, les mathématiques ont permis de construire, sur des bases saines, une véritable théorie générale des équations aux dérivées partielles.

La méthode des éléments finis

Au chapitre suivant, nous pousserons plus loin cette vision géométrique de l'analyse pour penser les fonctions comme des vecteurs dans un espace *euclidien* de dimension infinie. Mais pour terminer ce chapitre, je voudrais m'attarder sur une application *concrète* de la théorie des distributions à la simulation numérique de problèmes d'ingénierie.

Des équations différentielles pour modéliser le monde physique

Avant toute chose, il est important de bien comprendre que tous les phénomènes en physique classique peuvent être modélisés efficacement au travers d'équations différentielles – dont la plus célèbre est la deuxième loi de Newton.

Équation des cordes vibrantes Considérons par exemple une corde de violon, fixée par ses extrémités à un chevalet de longueur L. On peut repérer son évolution par la fonction d'altitude $y(x,t)$ au temps t, où l'abscisse x prend ses valeurs dans $[0,L]$. Quitte à changer de repère, on peut considérer que les conditions imposées aux bords du domaine s'écrivent

$$y(0,t) \;=\; 0 \;=\; y(L,t), \qquad \text{à tout instant } t. \tag{10.57}$$

Alors, selon la modélisation classique de D'Alembert (valide pour des cordes dont on peu négliger la raideur, et donc en particulier pour des oscillations de faible amplitude), on peut considérer que la corde est solution de l'équation

$$\frac{\partial^2 y}{\partial t^2} = v^2 \frac{\partial^2 y}{\partial x^2}, \tag{10.58}$$

où $v = \sqrt{T/\mu}$ est une vitesse caractéristique dépendant de la masse linéique μ de la corde et de la tension T exercée à ses extrémités. Notez que l'on privilégie ici une notation "avec dérivées directionnelles" plutôt qu'avec des "d" droits, suivant en cela les conventions des physiciens : $\frac{\partial y}{\partial t}(x,t_0)$ est la dérivée en t_0 de la fonction $t \mapsto y(x,t)$, tandis que $\frac{\partial y}{\partial x}(x_0,t)$ est la dérivée en x_0 de la fonction $x \mapsto y(x,t_0)$.

Cette équation se comprend très bien, à condition de savoir interpréter les symboles mis en jeu. Si $y(x,t)$ est un point courant de la corde à l'abscisse x au temps t, on voit que :

- $\frac{\partial y}{\partial t}$ est la *vitesse verticale* du point courant. Si elle est positive, c'est que la corde monte ; si elle est négative, c'est qu'elle redescend.
- $\frac{\partial^2 y}{\partial t^2}$ est l'*accélération verticale* du point courant, le taux de variation de la vitesse. Si elle est positive, c'est que la corde – au temps t et à la position x – a tendance à *remonter* : chute vers le bas *freinée*, ou remontée *accélérée*. À l'inverse, une accélération négative correspond à une attraction "vers le bas".
- $\frac{\partial y}{\partial x}$ est la direction de la *tangente* à la corde. Si elle est nulle, c'est que la corde est localement horizontale. Sinon, le signe de cette dérivée nous renseigne sur l'allure de la corde en x au temps t : montante s'il est positif, et descendante sinon.
- $\frac{\partial^2 y}{\partial x^2}$ est la *courbure* de la corde. Elle est négative si la corde est "courbée vers le bas" (comme $x \mapsto -x^2$), et positive si elle regarde vers le haut (comme $x \mapsto +x^2$).

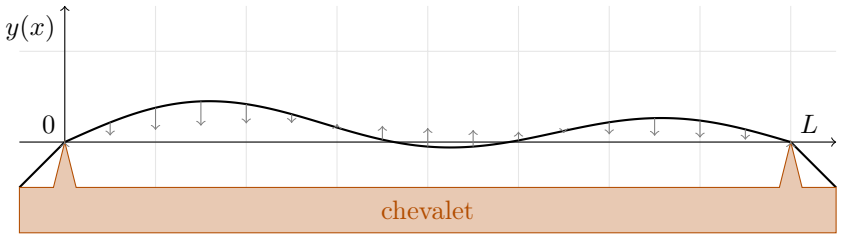


FIGURE 10.7 – Forces exercées le long d'une corde de violon, selon l'équation de D'Alembert.

L'équation des cordes vibrantes de D'Alembert peut donc s'énoncer directement en français :

« L'accélération d'un point de corde est proportionnelle à la courbure de celle-ci, avec une évolution en direction du raidissement. »

Un point où la courbe "regarde vers le haut" aura donc tendance à remonter, tandis qu'un point qui "regarde vers le bas" sera lui accéléré vers le sol. Avec une dynamique d'ordre 2 en temps (c'est l'accélération, en non la vitesse qui est proportionnelle à la courbure), on pourra montrer que les solutions de l'équation présentent un comportement *oscillant* autour d'une position d'équilibre $y_0(x) = 0$.

Généralisation aux membranes Le raisonnement peut être étendu des simples cordes aux membranes de tambour. Cette fois, l'altitude z dépendra des deux coordonnées x et y paramétrant la forme de la membrane fixée aux extrémités du domaine. La courbure moyenne de la membrane s'exprime alors au travers de l'opérateur *laplacien* bi-dimensionnel,

$$\Delta z = \frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2}, \tag{10.59}$$

qui est une expression indépendante du système de coordonnées orthonormées choisi. L'équation de D'Alembert pour les membranes s'exprime alors :

$$\frac{\partial^2 z}{\partial t^2} = v^2 \Delta z. \tag{10.60}$$

Équation de la chaleur Supposons donné un domaine Ω de dimension 1, 2 ou 3, muni de coordonnées (x), (x, y) ou (x, y, z). Pour plus de simplicité, on notera simplement r le point courant du domaine, indépendamment de la dimension. Alors, si $T(r, t)$ est la température au point r au temps t, on sait depuis les travaux de Fourier en 1822 que T est solution de l'équation :

$$\frac{\partial T}{\partial t} = D\Delta T + \frac{1}{\rho\, C_P} S, \tag{10.61}$$

où $S(r, t)$ est un terme de source (combustion, ...) apportant une chaleur à un point de capacité thermique $\rho(r) \cdot C_P(r)$ produit d'une masse volumique et d'une chaleur spécifique, tandis que $\Delta T(r, t)$ désigne le laplacien du champ de températures, diffusant dans le domaine selon une loi isotrope quantifiée par le coefficient D.

Dans le cas simplifié de la dimension 1, sans terme de source (système isolé) sur un domaine $\Omega = [0, L]$ (disons, une longueur L de fil de fer enrobé d'une gaine isolante), on trouve simplement l'équation

$$\frac{\partial T}{\partial t} = D\frac{\partial^2 T}{\partial x^2}, \tag{10.62}$$

que l'on peut utilement comparer à l'équation des cordes vibrantes 10.58. Typiquement, les conditions aux bord seront données par les températures aux extrémités : $T(0, t)$ et $T(L, t)$ fixées, imposées par le milieu extérieur.

Solutions analytique Ces deux équations liant une vitesse ou une accélération à une courbure (donnée par le laplacien) sont fondamentales en physique, étudiées avec soin depuis plus de deux-cents ans. Nous verrons au chapitre 3 comment décrire le comportement de leurs solutions en fonction des *harmoniques* du domaine étudié : la corde du violon, la membrane du tambour ou la plaque du radiateur.

À l'aide de formules explicites, on réussira (dans les cas où le domaine est bien homogène) à décrire l'évolution du système à partir d'une configuration initiale régulière.

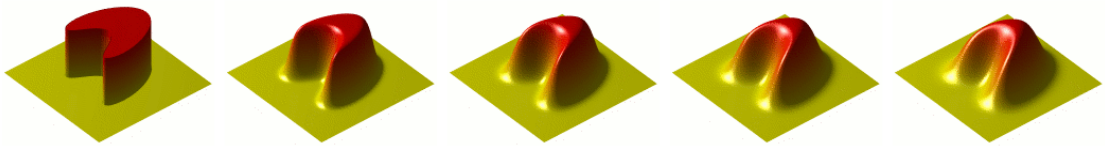


FIGURE 10.8 – Évolution d'une distribution de chaleur sur une plaque, suivant la loi de Fourier. L'évolution locale de la température est proportionnelle au laplacien du champ, équivalente ici à la courbure de la surface. Effet remarquable : la régularisation induite par l'équation, qui *lisse* la distribution initiale. Nous étudierons ce comportement au chapitre suivant. Image tirée de Wikipédia, par Oleg Alexandrov.

Solutions faibles Mais comment décrire l'évolution physique du système lorsque celui-ci n'est plus lisse ? Trois cas d'étude peuvent se présenter :

— Si l'équation est conservative (équation des cordes, qui conserve la régularité de l'état initial) ou régularisante (équation de la chaleur, qui l'augmente au fil du temps), on peut simplement chercher à comprendre ce qu'il se passe lorsque la condition initiale (y, z ou T en $t = 0$) est irrégulière (corde pincée, contact entre un fer chaud et un tissu frais).

— Il peut aussi s'agir d'étudier des termes de source irréguliers : sources *ponctuelles* de chaleur, masses ponctuelles en théorie de la gravitation, distributions surfaciques de charges en électromagnétisme... C'est d'ailleurs ce dernier cas qui inspira à Laurent Schwartz sa terminologie au début des années 50 : une *distribution* a vocation à représenter des mesures arbitraires de charges mais aussi des dipôles, des quadripôles, etc.

— Plus grave, certaines équations *font baisser la régularité du système* au fil du temps. C'est notamment le cas des équations de Navier-Stokes en mécanique des fluides : partant d'un écoulement lisse, il est possible d'arriver en temps fini à une situation irrégulière. C'est le phénomène d'émergence des *turbulences*, problème absolument fondamental en analyse avec les retombées que l'on imagine en aéronautique.

On s'en doute, il sera indispensable d'étendre l'étude des équations différentielles aux distributions. Heureusement, les opérateurs de dérivation sont parfaitement définis sur celles-ci : pour des fonctions à plusieurs variables, il suffit d'imiter la construction de la section 10.4. Donner un sens aux équations physiques sur les distributions générales ne pose donc aucun problème. L'essentiel de la recherche mathématique au sujet des équations différentielles se passera alors en deux temps :

1. Montrer que pour toute condition initiale raisonnable ("physique"), une solution existe à l'équation étudiée, au sens des distributions. Des résultats d'*unicité* peuvent également être démontrés.
2. Une fois la solution trouvée, le plus difficile reste à faire : montrer qu'elle est identifiable à une fonction, ou tout du moins à une configuration physiquement réaliste – énergie finie...

Les spécialistes d'un phénomène, d'une équation, chercheront donc à démontrer que pour toute configuration initiale "physique", une solution d'évolution est bien définie pour tout temps – en restant à chaque instant raisonnable. Si de tels résultats ont pu être obtenus pour les équations des cordes/surfaces vibrantes ou l'équation de la chaleur, le problème reste toujours ouvert pour les équations de Navier-Stokes ou d'Euler, en mécanique des fluides : fondamentalement non linéaires, elles résistent depuis deux-cents ans aux efforts de tous les théoriciens.

Résolution numérique d'équations aux dérivées partielles

Mais alors, comment travailler ? Les ingénieurs et physiciens ne peuvent attendre l'arrivée de résultats théoriques en se tournant les pouces... D'autant que l'espoir d'une résolution simple des problèmes de mécanique des fluides est bien mince : comment une formule *explicite* pourrait-elle décrire seule la foule de comportements possibles d'une lampée d'eau dans un verre ? Entre la chute d'un boulet et l'écume des vagues, il y a un monde de complexité.

La méthode de Galerkine Pour obtenir un problème numériquement résoluble, il est essentiel de restreindre l'espace des possibles. Attribuée à l'ingénieur russe Boris Galerkine (1871-1945), la très bonne idée sera de rechercher les solutions de nos équations sous la forme de combinaisons linéaires de fonctions élémentaires "crédibles".

Modélisation d'un problème concret Discutons par exemple du problème d'ingénierie par excellence : la construction d'un pont. Il s'agira pour nous de concevoir un *tablier* en acier, bois et béton simplement posé au dessus d'un ravin de longueur $L = 20\text{m}$, ouvrage que l'on modélisera naïvement par une fonction d'altitude $y(x)$ – où x prend ses valeurs dans l'intervalle $[0, L]$. En première approximation (c'est un cours de vulgarisation !), on verra notre pont comme un *fil* de masse linéique $\mu(x)$ et de raideur $k(x)$, propriétés qui dépendent du matériau utilisé au point x.

Mise en équation Par un raisonnement analogue à celui de D'Alembert pour les cordes vibrantes, on peut alors montrer que notre pont-jouet est solution de l'équation :

$$\mu \frac{\partial^2 y}{\partial t^2} = -\mu\, g + k \cdot \frac{\partial^2 y}{\partial x^2}, \tag{10.63}$$

où $g = 9.81\text{m}\cdot\text{s}^{-2}$ représente l'intensité du champ de gravitation, uniforme et orienté vers le bas. Si la dynamique du pont peut nous intéresser (Quelles sont les fréquences de résonance de notre ouvrage ? Risque-t-il de s'écrouler si une armée le traverse au pas ?), nous nous intéresserons ici uniquement à son profil au repos : étant données les informations de construction $\mu(x)$ et $k(x)$, sachant de plus que $y(x = 0) = 0 = y(x = L)$ (pont à l'horizontale), quel est le profil de notre ouvrage ? Présente-t-il une courbure, un affaissement trop important qui risquerait de le faire rompre sous son propre poids ? On l'aura compris, il s'agira ici d'étudier la solution du problème pratique :

$$k(x) \cdot \frac{\partial^2 y}{\partial x^2}(x) = \mu(x) \cdot g, \qquad \text{avec } y(0) = 0 = y(L). \tag{10.64}$$

Cette équation fort simple peut être résolue analytiquement : mais à quoi bon une formule compliquée pleine d'intégrales de k et μ ? Reposant sur la simplicité de l'équation, elle perdra toute valeur dès que l'on complexifiera le modèle. On cherche ici une méthode générale permettant de calculer, avec une précision arbitraire, le profil de notre tablier.

Premiers constats Commençons par interpréter notre équation : elle traduit simplement le fait que la *courbure* $\frac{\partial^2 y}{\partial x^2}(x)$ est d'autant plus importante au point x que le pont y est massif, et manque de raideur. Comme μ, k est g sont partout positives, la courbure l'est aussi : suivant notre modèle simplifié, le profil du pont est *convexe*. Ceci nous apprend par exemple que la fonction $y(t)$ est partout négative : en dessous de sa corde horizontale, le pont est bien en train de s'affaisser sous son propre poids.

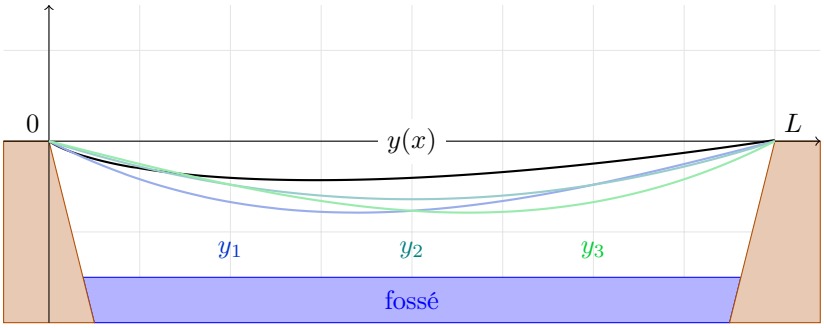


FIGURE 10.9 – Problème du pont, modélisé par l'équation (10.64). Les solutions y_1, y_2 et y_3 correspondent à des "profils crédibles" donnés équations (10.66-10.68). Plutôt que de se restreindre à une recherche dans un espace de dimension 3, un ingénieur moderne utilisera une famille de profils plus générique, comme celle illustrée Figure 10.10.

Un espace de fonctions de dimension finie Fort de cette information, Galerkine propose de chercher une solution à l'équation (10.64) de la forme

$$y(x) \;=\; \sum_{i=1}^{n} \lambda_i y_i(x) \;=\; \lambda_1 y_1(x) + \cdots + \lambda_n y_n(x), \tag{10.65}$$

où les $y_i(x)$ sont des profils "crédibles" fixés a priori et où les λ_i sont des coefficients réels qui vont nous permettre d'affiner notre recherche jusqu'à trouver une "bonne" solution. En 1915, Galerkine proposerait par exemple de nous restreindre à une famille de trois fonctions,

$$y_1(x) = x \cdot (x - L) \cdot (2L - x) \tag{10.66}$$
$$y_2(x) = x \cdot (x - L) \tag{10.67}$$
$$y_3(x) = x \cdot (x - L) \cdot (L + x) \tag{10.68}$$

qui sont toutes convexes sur $[0, L]$, vérifient les conditions aux bords et ont des minimums distincts.

Formulation faible avec espace réduit de fonctions tests Bien sûr, impossible d'espérer que des coefficients λ_1, λ_2 et λ_3 nous donnent une solution exacte au problème (10.64) : la véritable solution n'a aucune raison d'être polynomiale. Mais si nos *fonctions de base* y_i sont bien choisies, on devrait tout de même pouvoir trouver une solution "satisfaisante".

Point clé En quel sens ? Comment affaiblir le fait d'être "solution d'une équation différentielle" ? Eh bien, le formalisme des distributions va nous donner une manière bien posée et robuste de penser ce problème. Là où une "vraie" solution faible de notre équation devrait vérifier

$$\forall\, \varphi \in D, \left\langle k(x) \cdot \frac{\partial^2 y}{\partial x^2}(x), \varphi \right\rangle \;=\; \langle \mu(x) \cdot g, \varphi \rangle\,, \tag{10.69}$$

une solution "au sens de Galerkine" devra vérifier cette égalité pour les seules fonctions test **qui sont elles-mêmes combinaisons linéaires des** y_i. À vrai dire, par linéarité de l'équation en φ, il suffira même de vérifier cette identité pour les seules fonctions tests $\varphi = y_1, \ldots, y_n$.

Résoudre l'équation du pont au sens de Galerkine, c'est donc trouver des coefficients optimaux λ_j tels que

$$\forall\ i \in [\![1,n]\!]\,,\ \left\langle k(x)\cdot\frac{\partial^2 \sum_j \lambda_j y_j}{\partial x^2}(x), y_i\right\rangle\ =\ \langle \mu(x)\cdot g, y_i\rangle\,, \tag{10.70}$$

i.e., en développant le terme de droite par linéarité de la dérivation et du crochet de dualité :

$$\forall\ i \in [\![1,n]\!]\,,\ \sum_j \lambda_j \left\langle k(x)\cdot\frac{\partial^2 y_j}{\partial x^2}(x), y_i\right\rangle\ =\ \langle \mu(x)\cdot g, y_i\rangle\,. \tag{10.71}$$

Si on note

$$a_{ij} = \left\langle k(x)\cdot\frac{\partial^2 y_j}{\partial x^2}(x), y_i\right\rangle \qquad \text{et} \qquad b_i = \langle \mu(x)\cdot g, y_i\rangle \tag{10.72}$$

qui sont des coefficients réels explicites, il s'agit donc de trouver une solution au système linéaire

$$\left\{\begin{array}{ccccccccc} a_{11}\lambda_1 & + & a_{12}\lambda_2 & + & \dots & + & a_{1n}\lambda_n & = & b_1 \\ a_{21}\lambda_1 & + & a_{22}\lambda_2 & + & \dots & + & a_{2n}\lambda_n & = & b_2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ a_{n1}\lambda_1 & + & a_{n2}\lambda_2 & + & \dots & + & a_{nn}\lambda_n & = & b_n \end{array}\right.\,, \tag{10.73}$$

que l'on écrit simplement sous forme matricielle : $A\lambda = B$. Avec n équations pour n inconnues, ce système admet généralement une unique solution, qui caractérisera le meilleur candidat à la résolution de l'équation différentielle dans l'espace engendré par les y_i. Pour simuler de manière approchée un phénomène physique, il suffira donc de choisir une base d'approximation (y_i), de construire les matrices A et B par un calcul direct, puis de résoudre le système associé par un algorithme bien connu.

Pertinence de la formulation faible Pour qui pense les fonctions comme des *graphes*, il est malaisé de définir ce qu'est une solution "approchée" f à une équation différentielle du type

$$(E) : \forall\ x \in \Omega,\ A(f)(x) = 0. \tag{10.74}$$

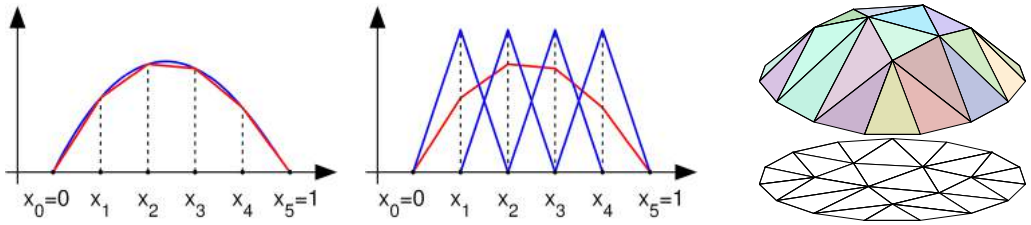
Si, dans un premier temps, on demande que f vérifie (E) *exactement* sur un *sous-ensemble* du domaine Ω – disons, sur une discrétisation fixée à l'avance, un point tous les 50cm – alors impossible de garantir son comportement en dehors. C'est le grand défaut des polynômes interpolateurs de Lagrange, qui peuvent faire exploser la fonction interpolée en dehors des points de contrôle.

À l'opposé, on pourrait demander à ce que f vérifie (E) de manière approchée, uniformément sur le domaine : que le maximum sur Ω de la fonction $x \mapsto |A(f)(x)|$ soit aussi petit que possible ; que la fonction $A(f)$, si elle ne peut être nulle, s'en approche du moins du mieux qu'elle le peut. Mais une telle définition implicite n'est pas d'une grande aide en pratique...

En équilibrant l'approximation sur l'espace de recherche et sur celui des contraintes, Galerkine nous a tiré d'un mauvais pas : ses solutions faibles seront aussi bien conditionnées que le permet l'espace de recherche engendré par les y_i, mais n'en resteront pas moins des solutions "exactes", issues d'une égalité et non d'un problème de minimisation mal posé.

Méthode des éléments finis En un siècle, grâce aux machines de calcul électroniques, la résolution de systèmes linéaires a fait des pas de géant : utiliser la méthode de Galerkine avec des milliers de fonctions arbitraires y_i n'est plus un problème. Il n'y a donc plus besoin de les prendre dans un ensemble de solutions "crédibles" : seule compte l'expressivité de l'espace engendré. On pourra par exemple se contenter de fonctions y_i "en chapeau pointu", présentées Figure 10.10, qui permettent d'engendrer l'espace des fonctions affines par morceaux associées à une discrétisation donnée et qui cumulent les avantages. Construites à partir de la valeur absolue, leurs dérivées sont connues. Très localisées, elles garantissent au système (10.73) une structure *parcimonieuse* garantie d'efficacité algorithmique. Simplement définies, elles se généralisent sans difficultés aux espaces de dimensions 2, 3 ou 4. Seuls compteront alors le choix d'une discrétisation adaptée au problème, et l'efficacité du schéma de résolution numérique.

On pourra facilement monter jusqu'à des valeurs de n de l'ordre du million et plus, pour des problèmes industriels complexes résolus sur des centres de calcul. Simuler des tests en soufflerie, des crashs de voitures en s'épargnant le fastidieux travail d'acquisition des données. Ce grand succès de l'ingénierie du XX^{e} siècle, l'application de la méthode de Galerkine a une famille de fonctions y_i élémentaires adaptées, c'est la *méthode des éléments finis*.

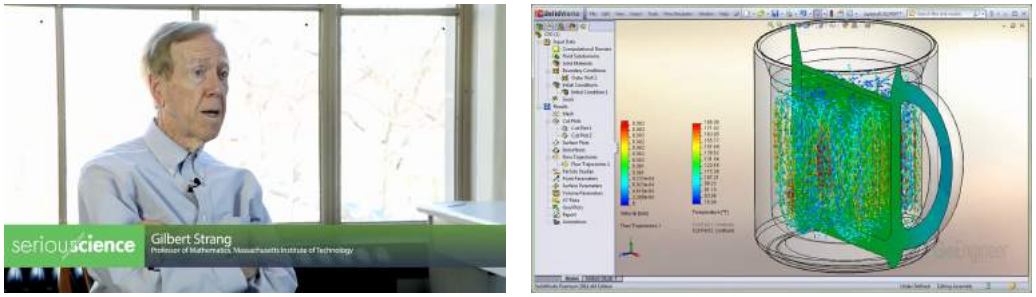


(a) Approximation affine par morceaux d'une courbe lisse...

(b) Combinaison linéaire d'une famille en "chapeaux pointus".

(c) En dimension 2 et plus, des familles analogues existent.

FIGURE 10.10 – Des familles de fonctions y_i affines par morceaux, localisées et liées à une discrétisation du domaine permettent d'approcher n'importe quel profil. Images tirées de Wikipédia.



(a) *Finite element method - Gilbert Strang*, de la chaîne "Serious Science" ; `www.youtube.com/watch?v=WwgrAH-IMOk`.

(b) *Flow Simulation - Transient Natural Convection*, QuickTips video presented by Tony Botting of GoEngineer ; `www.youtube.com/watch?v=hSEYBkgcmhA`.

FIGURE 10.11 – En cours, nous regarderons deux vidéos disponibles sur YouTube : une interview de Gilbert Strang, auteur de l'excellent manuel *An Introduction to Applied Mathematics*, et une démonsration du logiciel Solidworks, outil de référence en ingénierie édité par Dassault Systèmes. Ce sera l'occasion de faire le lien entre les mathématiques présentées dans ce chapitre et leurs applications industrielles, d'une importance capitale. Si le logiciel Solidworks se révèle absolument "bluffant", on trouvera au cœur de son moteur de rendu physique la méthode des éléments finis : rien de magique, donc ; seulement des mathématiques bien comprises.

Appendices

Annexe A

Arithmétique

Annexe B

Éléments de topologie : continuité, limites et points fixes

La notion de limite

Continuité

Théorème des valeurs intermédiaires

Compacité

Deux illustrations : les théorèmes de point fixe

Le théorème de Brouwer

Le théorème de Cauchy-Lipschitz