Notes et commentaires au sujet des conférences de S. Mallat du Collège de France (2023)

Jean-Eric Campagne

Notes et commentaires au sujet des conférences de S. Mallat du Collège de France (2023)

Modèles multi-échelles et réseaux de neurones convolutifs

J.E Campagne *

Janv. 2023; rév. 10 mars 2024

*Si vous avez des remarques/suggestions veuillez les adresser à `jeaneric DOT campagne AT gmail DOT com`

Table des matières

1. Avant-propos

***Avertissement**: Dans la suite vous trouverez mes notes au style libre prises au fil de l'eau et remises en forme avec quelques commentaires ("ndje" ou bien sections dédiées). Il est clair que des erreurs peuvent s'être glissées et je m'en excuse par avance. Vous pouvez utiliser l'adresse mail donnée en page de garde pour me les adresser. Je vous souhaite une bonne lecture.*

Veuillez noter que le site web du Collège de France a été remanié, vous trouverez toutes les vidéos des cours, des séminaires ainsi que les notes de cours non seulement de cette année mais aussi des années précédentes [1].

Je tiens à remercier l'ensemble de l'équipe du Collège de France qui réalise et monte les vidéos sans lesquelles l'édition de ces notes serait rendue moins confortable.

Notez également que S. Mallat [2] donne en libre accès des chapitres de son livre *"A Wavelet Tour of Signal Processing"*, 3ème edition. ainsi que d'autres matériels sur son site de l'ENS.

Cette année 2023 c'est la sixième du cycle de la chaire de la Science des Données de S. Mallat, le thème en est: **Modélisation, Information et Physique Statistique**.

Dans le cadre de cette année deux livres sont recommandés: **"Elements of Information Theory"** de Thomas Cover et Joy Thomas [3], et **"Information, Physics and Computation"** de Marc Mézard et Andrea Montanari [4]

NDJE. Je profite pour dire que j'ai mis en œuvre un repository GitHub pour quelques applications numériques illustrant le cours. https://colab.research.google.com/github/jecampagne/cours_mallat_cdf. Pour 2023, les notebooks pourront directement être exécutés sur Google Colab. La migration des nbs de 2022 est également prévue.

Enfin, en ce début de 2023, la pandémie de Covid-19 est toujours en sous-jacent mais nous avons des vaccins efficaces, cependant nous sommes toujours inquiets par la

1. https://www.college-de-france.fr/chaire/stephane-mallat-sciences-des-donnees-chaire-statutaire/events
2. https://www.di.ens.fr/~mallat/CoursCollege.html
3. https://ia801400.us.archive.org/30/items/ElementsOfInformationTheory2ndEd/Wiley_-_2006_-_Elements_of_Information_Theory_2nd_Ed.pdf
4. https://cds.cern.ch/record/1166773/files/9780198570837_TOC.pdf

Guerre en Ukraine déclenchée il y a presque 1 an par la Russie. Gageons que ces cours viennent nous distraire de cette atmosphère pesante.

2. Séance du 18 Janv.

2.1 Introduction

Le cours de cette année, nous dit Stéphane Mallat, tournera autour de trois mots: la notion de **Modèles** de données en grande dimension ($x \in \mathbb{R}^d$), la théorie de l'**Information** qui était déjà au cœur du cours de 2022, et finalement la **Physique Statistique** qui nous occupera cette année. La question qui guidera notre cheminement est la suivante: ***comment penser les problèmes et modéliser les données en grande dimension***? En fait, il n'est pas intuitif de penser en grande dimension et c'est là où la Physique Statistique va nous aider.

Coté "données" nous nous intéresserons aux traitements d'images, du son, du texte, aux mesures physiques, chimie... Coté "applications", une fois que l'on a un modèle des données, on peut faire de l'inférence des paramètres, on peut vouloir enlever du bruit, résoudre des problèmes inverses, et envisager des problèmes de classification/régression. Ces notions ont déjà été abordées dans les années précédentes. La nouveauté sera le point de vue envisagé avec en sous-jacent des concepts de la Physique Statistique afin de construire un modèle et en voir les conséquences.

2.2 La Physique Statistique: pourquoi?

Rappelons que pendant un siècle en gros la Physique Statistique était la seule science de la grande dimension avant l'avènement des problématiques d'apprentissage à la manière dont on l'entend actuellement. Au passage, notons que Josiah Willard Gibbs (1839-1903) écrit en 1901 un ouvrage [5] établissant un pont solide entre la Mécanique Statistique et la Thermodynamique, et généralise l'interprétation statistique de l'**entropie** d'un système.

5. J. W. Gibbs, *Elementary Principles in Statistical Mechanics developped with especial reference to the Rational Foundation of Thermodynamics*, Yale Univ. publié en Mars 1902.

Laquelle entropie est au cœur de la théorie de Shannon développée dans les années 1940 et que nous avons étudiée en 2022.

Une *mole* de quelque chose contient par définition $6\,10^{23}$ entités (cf. le nombre d'Avogadro), ce qui fixe la dimension [6] du système d. On est clairement dans le cadre de la grande dimension. Afin d'appréhender les phénomènes macroscopiques étudiés par la Physique, un certain nombre de notions ont été dégagées au fil du temps comme: l'**énergie** qui peut se conserver, les notions de **forces** appliquées sur le système, la notion d'**équilibre**, puis avec la Physique Statistique apparait la notion de **probabilité** et surtout la notion d'**entropie**, enfin avec les théories qui sont à la base du Modèle Standard de la Physique des Particules et de la Cosmologie mais également en lien avec la Physique Statistique des Transitions de Phases va apparaitre des notions centrales telles que les **symétries**, les **interactions** et les **échelles**.

L'idée est donc de montrer que les notions dégagées dans le champ de la Physique Statistique vont naturellement apparaitre dans l'Apprentissage où la *modélisation de données* joue également un rôle central. Dans ce contexte, la **Théorie de l'Information** va jouer le rôle d'un médiateur, selon la formule de Stéphane Mallat. Tout d'abord, elle a donné un *point de vue mathématique* aux notions citées ci-dessus, c'est-à-dire d'en faire des notions détachées du contexte où elles sont apparues, et ce faisant ces nouvelles notions seront applicables dans d'autres situations. En premier lieu bien entendu nous retrouvons l'**entropie** associée aux probabilités, mais aussi **les ensembles typiques** avec les applications au codage, aux transmissions (voir cours de 2022). Ainsi, les notions abordées dans le cours de 2023 sont schématisées sur la figure 1.

2.3 Le point de vue (rapide) de Physique Statistique

Disons qu'à gros traits jusqu'à la fin du 19e siècle les champs de la Physique tels que la Mécanique du solide, celle des milieux continus, l'Electro-magnétisme, l'Optique, la Thermodynamique, la Chimie, traitent de propriétés macroscopiques, avec leurs lots de lois et Principes Fondamentaux. A la fin du 19e début du 20e siècle, il y a un changement qui s'opère avec le point de vue de la Mécanique Statistique (qui devient Physique

6. nb. l'espace de phase au sens classique est de dimension $6d$ en l'occurrence.

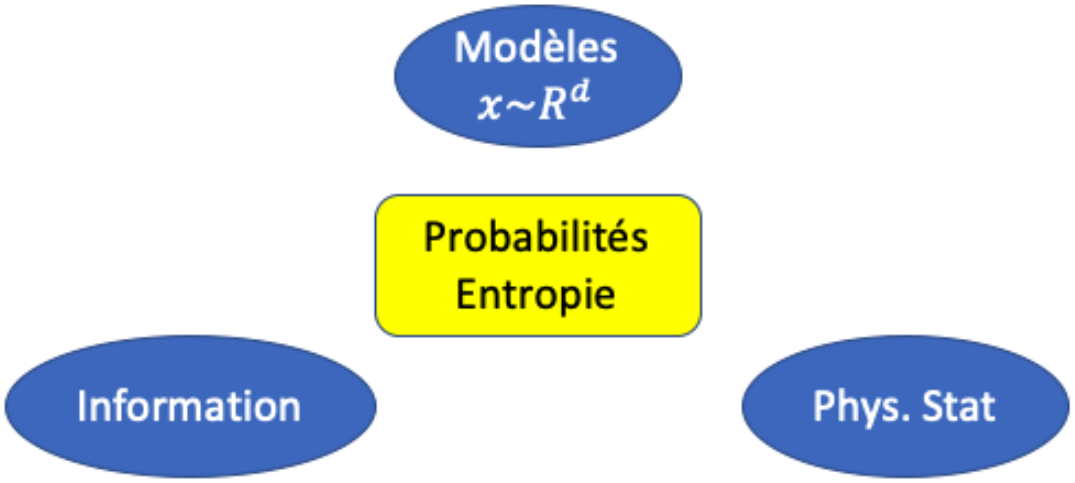


FIGURE 1 – Les différentes notions abordées dans le cours de 2023.

Statistique). On peut dire que cela commence par une hypothèse que les propriétés macroscopiques sont en fait le résultat de collisions à l'échelle microscopique entre composants (ex. atomes, particules) et même si la réalité expérimentale fait défaut à l'époque, J. Clerk Maxwell (1831-79) développe la théorie de la distribution des vitesses dans les gaz qui est généralisée en 1896 par Ludwig Boltzmann (1844-1906) qui interprète l'entropie [7] selon la fameuse formule "S = k log W" qu'il fait graver sur sa tombe. Puis comme déjà dit, J. W. Gibbs établit en 1901 un pont entre Mécanique Statistique et Thermodynamique.

Ce faisant tous les champs disciplinaires cités vont être réorganisés selon un axe des échelles allant de la taille des particules élémentaires (1 fermi, 10^{-15}m), la taille des atomes (1 angström, 10^{-10}m) jusqu'aux échelles astrophysiques (1 ly, année lumière, 10^{16}m) et cosmologiques (Gly), en passant par les tailles "humaines". Les variables/grandeurs décrivant les phénomènes en Mécanique du Solide, Chimie etc sont somme toute similaires d'un point de vue mathématique. Le problème est que si on a au niveau microscopique des équations bien établies - soit en classique celles de Newton, Maxwell, etc, soit en quantique celles de Schrödinger, Dirac, etc - comment en déduire les propriétés aux échelles macroscopiques?

7. Le premier concept d'entropie (1854) est dû à Rudolf Clausius (1822-88).

L'idée principale de Maxwell, Boltzmann et Gibbs est qu'il est sans doute vain de vouloir étudier l'évolution d'un système particulier à N molécules surtout dans le contexte de la Mécanique Hamiltonienne, mais il serait considéré comme plus opportun d'étudier un ensemble de systèmes pour établir les lois qui les caractérisent dans en un premier temps à l'équilibre thermodynamique, d'où l'introduction de la statistique des systèmes. Le point non évident que Gibbs finit de démontrer, c'est que les grandeurs et fonctions édifiées en Thermodynamique trouvent un pendant dans ce cadre statistique qui en retour enrichit la description (ex. potentiel chimique introduit par Gibbs). Il est remarquable que les résultats de Mécanique Quantique n'ont pas chamboulé ce cadre théorique et les Statistiques de particules indiscernables de Bose-Einstein (1924-25) et Fermi-Dirac (1926) y trouvent naturellement leur place. La Physique Statistique est en mesure de comprendre l'apparition des transitions de phase.

Donc, la Physique Statistique permet de comprendre l'émergence de propriété quand on passe du *microscopique* au *macroscopique*. Or, cette évolution est très similaire à celle que l'on observe en Science des Données. Disons là aussi à grands traits, que jusqu'à 2015 environ, il y avait des domaines distincts comme la Vision par ordinateur (imagerie), le Son (avec des sous domaines comme la parole, la musique, etc), l'analyse du Texte et du Langage, et d'autres comme l'Imagerie Médicale, voire l'analyse de données physiques, etc. Ceci dit il y avait le Traitement du Signal qui était une théorie commune sous-jacente, mais les types de données, les types d'algorithmes étaient différents, et *a fortiori* les communautés de recherche étaient largement cloisonnées. Mais comme on le sait maintenant, force est de constater que les réseaux de neurones profonds ont changé l'ensemble de ces disciplines, au grand dam des différents experts car leur savoir s'en est trouvé concurrencé voire mis à l'écart par des utilisateurs de Machine Learning. Certes déprimant pour certains, il faut reconnaitre qu'il a émergé de cette décennie qui peut être qualifiée celle "des pionniers", une structure ***de réseaux en cascades de convolution-pooling-non linéarité*** comme celle schématisée sur la figure 2 qui fonctionne ***d'une manière universelle dans tous les domaines précités***. Noter qu'à l'entrée du réseau de neurones, les filtres convolutionnels scannent des *petites échelles* de l'image [8], et plus l'on va profond dans le réseau, plus les filtres couvrent des zones de plus en plus grandes tailles, pour délivrer *in fine* une information globale ex. catégories chien/chat, mot suivant d'un texte, etc. On

8. On illustre le propos avec des images car soucis d'illustration mais le schéma peut être étendu aux autres types de données.

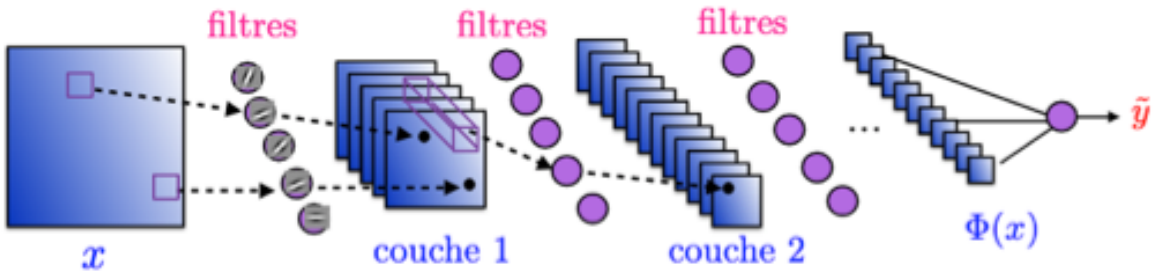


FIGURE 2 – Schématisation d'un réseau profond convolutionnel typique.

passe donc du microscopique au macroscopique. Et comme en Physique, on a un nouvel axe d'organisation qui couvre toutes les disciplines avec la notion de ***multi-échelles*** très complexes avec les non linéarités.

Nous savons qu'au fil de ces dix dernières années ont émergé des architectures de réseaux de plus en plus complexes avec des succès d'autant meilleurs, et malgré tout il n'y a pas beaucoup de théorie mathématique permettant d'expliquer ces succès. C'est la différence majeure avec la Physique Statistique qui élaborée durant un siècle apporte un support théorique bien solide. S'il n'est pas étonnant que les concepts quelle a dégagé puissent être utiles en Apprentissage, cependant il ne va pas s'agir de faire une sorte de "Physique Statistique Appliquée". En effet, ***les réseaux de neurones profonds sont capables de résoudre des problèmes bien plus sophistiqués que ceux étudiés en Physique Statistique, mais cette dernière peut nous servir comme guide conceptuel comme nous allons le voir cette année, le but étant de comprendre et d'établir une théorie des réseaux de neurones.***

2.4 Les concepts de la Physique Statistique

Comment les concepts de Physique Statistique ont émergé au fil du temps? Bien entendu on ne peut être exhaustif dans cette quête historique et d'ailleurs sans doute l'exhaustivité brouillerait le message. Donc la suite est une mise en perspective pour nous servir de guide. Ceci étant dit, on peut faire débuter le processus théorique par la

Thermodynamique avec ses deux Principes [9].

Le premier Principe stipule que pour un système à l'équilibre, lors d'une transformation, on peut définir une fonction (***énergie interne***) qui a trois propriétés: la fonction d'état ne dépend que des états initiaux et finaux de la transformation, elle est extensive, et elle se conserve pour un système isolé. La ***conservation*** de l'énergie est en fait une conséquence de l'***invariance*** des lois de la Physique par rapport au temps. Notons que cette *relation entre conservation et invariance* est un duo extrêmement important, tout à fait fondamental et la base de création de modèles en Physique des particules et Gravitation, dont un théorème fameux d'Emmy Noether (1882-1935), mathématicienne allemande, en est le sujet.

Le second Principe réinterprété de nombreuses fois dont l'origine est le ***Principe de Carnot*** (établi en 1824 par Nicolas Léonard Sadi Carnot 1796-1832), porte sur la notion d'***irréversibilité*** en stipulant "qu'au cours d'une transformation, il y a une partie de l'énergie qui est irréversiblement perdue en chaleur." Cette notion a été reliée à celle d'***entropie*** (voir Sec. 2.3) en formulant le Principe selon lequel: "*toute transformation d'un système thermodynamique s'effectue avec augmentation de l'entropie globale incluant l'entropie du système et du milieu extérieur. On dit alors qu'il y a création d'entropie.*" Il est remarquable qu'entre les travaux de R. Clausius (1854) et ceux de L. Boltzmann (1872) une vingtaine d'années ont fixé le cadre théorique de l'entropie sous jacente de cette notion d'irréversibilité. D'un point de vue pratique dans la suite du cours nous noterons H l'entropie en accord avec la notation de Cl. Shannon [10]. Le principe de Clausius peut se formaliser dans l'inégalité suivante où δQ est la variation de chaleur, T la température et δH la variation d'entropie

$$\delta H_{système} \geq \frac{\delta Q_{échangée}}{T} \tag{1}$$

Cependant, cette relation fut-elle établie expérimentalement par Clausius, n'est pas issue d'une théorie mathématique à proprement parler. En quelque sorte, toute chose égale par

9. nb. la notion de *Principe* surtout à l'époque de leurs élaborations est à prendre au sens de moyen d'*organiser* les savoirs.

10. nb. au passage pour les curieux voici le texte original des cours de Boltzmann traduis en français et préfacés par Léon Brillouin: https://gallica.bnf.fr/ark:/12148/bpt6k98134700/f1.item.texteImage; Boltzmann utilise la lettre H et il s'agit de l'opposée de l'entropie; H en thermodynamique est la notation de l'enthalpie, et en mécanique H peut également être la notation de l'Hamiltonien du système, donc attention au contexte.

ailleurs, en Thermodynamique on serait dans la même situation qu'actuellement avec les réseaux de neurones, si L. Boltzmann et J. W. Gibbs n'étaient pas passés par là.

L. Boltzmann va établir la notion d'entropie comme dit plus avant en continuant les travaux de J. C. Maxwell sur la *cinétique des gaz*, appelées maintenant théorie de Maxwell-Boltzmann. C'est Maxwell qui introduit les probabilités dans ce domaine [11], et Boltzmann fait une démonstration de l'unicité de la distribution de Maxwell pour des gaz *sans organisation d'ensemble* [12] (*parfait*) même si L. Brillouin est sceptique sur la validité de la démonstration bien que reconnaissant la portée de la fonction H de Boltzmann [13]. Donc, la distribution de la position r_i d'une molécule i du gaz dépendant de sa vitesse v_i évolue dans le temps $p(r_i, v_i, t)$ selon une équation différentielle (Equation de Boltzmann) qui atteint un point d'équilibre stationnaire, et la fonction [14] $H = -\sum_i \int_v p_i \log p_i dv_i$ va déprendre du nombre de configurations possibles du système (W) et donne la relation de proportionalité suivante [15]

$$H = k_B \log W \tag{2}$$

(nb. k_B est une notation moderne de la constante de Boltzmann). Boltzmann nous dit en substance que le système va évoluer vers un état où toutes les configurations (*micro-états*) sont équivalentes, ce que l'on établit en ***Principe Fondamental d'équiprobabilité***. Ces ensembles de configurations (*micro-canoniques*), nous le verrons par la suite, correspondent à un certain type de modèles pour lesquels la probabilité est totalement uniforme et comme le système est isolé, son ***énergie est fixe*** (une constante). L'irréversibilité est comprise comme l'évolution de l'entropie qui ne fait qu'augmenter avec le temps.

Le problème avec le postulat d'équiprobabilité est qu'en général le système en question est rarement totalement isolé et donc "l'univers" est composé du système auquel il

11. Voir Robert E Robson et al 2017 Eur. J. Phys. 38 065103: *Great moments in kinetic theory: 150 years of Maxwell's (other) equations.* https://iopscience.iop.org/article/10.1088/1361-6404/aa87d4/pdf

12. *molar-ungeordnet*

13. Voici un extrait de la postface du texte original mentionné en note de bas de Page 12): *"... Outre que je suis inquiet sur la légitimité de la division adoptée pour tout l'espace, qui est ici infini, je ne vois aucune raison qui permette de proposer une loi de probabilité de la vitesse d'une molécule isolée. Passons outre, et adoptons la proposition de M Boltzmann: dans cet ordre d'idées, la distribution des vitesses de Maxwell serait la plus probable, parce que c'est elle qui comporte le plus grand nombre possible de permutations..."*.

14. nb. ici le signe "moderne" de l'entropie.

15. L'expression original de Boltzmann considère est sa fonction $-H$ de son théorème H, c'est Max Planck qui écrit la formule $S = k_B \log W$ dans une conférence en 1923 sur la théorie du rayonnement thermique (*Wärmestrahlung*). •

faut ajouter son environnement. C'est pour cette raison que *Gibbs étudie des systèmes en présence de réservoir*, ce qui l'amène à revisiter le concept d'entropie. Il approfondit la relation entre probabilité et entropie, où $H = -\sum_c p_c \log p_c$ (ici avec des états discrets de configurations), et la notion d'***ensembles micro-canoniques, macro-canoniques***. Nous avons vu que pour un système micro-canonique l'énergie est constante, tandis que pour un système macro-canonique, l'énergie fluctue et seule la moyenne est constante. Le réservoir (beaucoup plus grand) sert à fixer la température par exemple. Avec les développements de Gibbs, apparaissent des notions comme la ***transformation de Legendre***, les ***variables duales*** (température-énergie, pression-volume, etc). L'outil de Gibbs est la ***fonction de partition*** Z à partir de laquelle il va retrouver les lois de la Thermodynamique.

Avec les notions d'ensembles micro/macro-canoniques en apprentissage vont correspondre deux types de modèles de données différents (nous y reviendrons) tant que la dimension d du système est finie (c'est-à-dire à nombre d'atomes/molécules/entités fini) mais deviennent équivalents à la limite thermodynamique c'est-à-dire quand $d \to \infty$. Donc, si numériquement les deux modélisations donnent des résultats équivalents, les algorithmes sont différents comme dans l'exemple de la génération de données.

2.5 Petit détour historique

NDJE. Ce qui suit est une version bien entendu partielle et on ne peut être exhaustif en la matière dans ce cours. J'ai ajouté quelques remarques personnelles au déroulé présenté par S. Mallat, en espérant que le lecteur n'en prendra pas ombrage.

Notons que les points de vue de Boltzmann et Gibbs donc à la fin du 19e siècle, se basent sur l'***existence de corpuscules élémentaires***, sans qu'aucune preuve irréfutable expérimentale ne vienne étayer cette hypothèse. D'ailleurs d'autres physiciens non moins célèbres étaient en faveur d'une matière continue comme Marcelin Berthelot (1827-1907) [16]. Mais comme on le sait les preuves vont voir le jour au tournant du 20e siècle. On peut citer brièvement l'enchainement suivant. Wilhelm Röntgen en 1895 découvre les ***rayons X*** (Nobel de Physique en 1901), puis Max von Laue (1879-1960) découvre en 1912 que les rayons X sont diffractés par ***les cristaux*** (Nobel de Physique en 1914) ce qui donne à

16. ce qui a eu des conséquences fâcheuses pour le développement de la Chimie française, voir dans l'Encyclopédie Universalis l'article https://www.universalis.fr/encyclopedie/theorie-atomique/2-la-resistance-de-berthelot/.

première vision de l'aspect "atomique" de la matière, enfin si l'on peut dire en 1909 Hans Geiger (1882-1945), Ernest Marsden (1889-1970) et Ernest Rutherford (1871-1937) (Nobel de Chimie en 1908) établissent expérimentalement ***le modèle planétaire de l'atome*** avec une charge positive centrale réfutant par la même le modèle d'une charge positive diffuse de J.J Thomson (1856-1940). On a alors une confirmation claire de la nature corpusculaire de la matière déjà à plusieurs échelles: atome et noyau.

Ce faisant le physicien John William Strutt Rayleigh [17] (1842-1919) associé au mathématicien physicien James Jeans (1877-1946) utilise la Mécanique Statistique afin d'établir en 1900 la loi, connue sous le nom de ***loi de Rayleigh-Jeans***, qui exprime la répartition de l'énergie rayonnée par ***le corps noir*** en fonction de la longueur d'onde, valable pour les grandes longueurs d'onde, que Max Planck [18] (1858-1947) complétera la même année en utilisant ***l'hypothèse des quanta*** (Nobel de Physique en 1918).

Du coté de la Physique Statistique à l'époque de Gibbs, il restait le mystère de la description dans ce schéma théorique des ***changements d'états*** ou ***transitions de phases*** (passages solide/liquide/vapeur, changement ferro/para-magnétisme, condensation quantique, etc) lors de franchissement de seuil en température. Les postulats de la théorie pouvaient ne pas être les bons... Notez au passage que cela soit la fonction de partition Z ou l'énergie libre de Gibbs sont des fonctions analytiques de la température alors pourquoi doit-il y avoir des singularités? Il a fallu des travaux dont ceux de Hendrick Kramers (1894-1952) et Gregory Wannier (1911-83) puis finalement Lars Onsager (1903-76) et Bruria Kaufman (1918-2010) pour montrer que les transitions de phase se manifestent dans le cadre de la Mécanique Statistique de Gibbs ce qui fût un tournant décisif. Le problème qu'Onsager a résolu exactement en 1944 est le depuis fameux modèle d'Ising 2D: ce modèle de spins en interaction a été introduit par Wilhelm Lenz (1888- 1957) en 1920 et son étudiant Ernest Ising (1900-98) l'avait résolu en 1D uniquement et n'avait pu trouver de transition de phase. La résolution exacte de Onsager a permis d'en comprendre le sens et d'impulser l'étude des exposants critiques et le développement en Mécanique Statistique de la ***Théorie des Equations du Groupe de Renormalisation*** (RGE). Notez au passage que le mathématicien Hugo Duminil-Copin (1985-) a reçu la Médaille Fields 2022 pour ses

17. Il reçoit le Nobel de Physique en 1904 pour des études sur les gaz ayant amené à la découverte de l'Argon, travaux réalisés en commun avec William Ramsay (1852-1916) qui de son coté reçoit la même année le Nobel de Chimie également pour la découverte de l'Argon.

18. Initialement Planck était en faveur d'une matière continue rejettant la théorie statistique, mais il se rend à l'évidence et adopte le point de vue dit atomiste à la suite des résultats expérimentaux.

travaux entre autres sur le modèle d'Ising en dimension 3 et 4 [19], et que Giorgio Parisi [20] (1948-) a reçu le Nobel de Physique 2021 pour ses travaux sur les systèmes désordonnés et l'étude des brisure de symétrie des répliques. Donc, la recherche en Physique Statistique est toujours très active.

Ce qui est particulièrement intéressant d'un point de vue historique c'est que Physique Statistique et Physique des Particules ont échangé beaucoup de concepts. Le mécanisme de "Higgs" de 1964 [21], en est un exemple où Higgs (Phys. des Particules) utilise la brisure spontanée du boson scalaire de Brout-Englert (Phys. Stat.) pour donner la masse aux bosons W et Z médiateurs de l'interaction faible. La RGE a été initiée en Théorie des Champs en Physique de Particules en 1954 par Murray Gell-Mann (1929-2019) et Francis E. Low (1921-2007) dans le cadre de la QED (Quantum Electrodynamics), puis elle fût généralisée par Curtis Callan et Kurt Symanzik (1923-83) par l'établissement les équations de Callan–Symanzik en 1970. Les développements sur les transitions de phases datent Leo Philip Kadanoff (1937-) puis du mémoire du Ph. D de Kenneth G. Wilson (1936-2013) obtenu sous la direction de Gell-Mann en 1961. Wilson fait le lien avec les développements en Théorie des Champs et développe la théorie des exposants critiques en lien avec les transitions de phases qui sera un thème de choix du domaine dans les années 70s comme le fameux "Les Houches Session XXVIII (1975): Methods in Field Theory" avec des contributions exceptionnelles.

Le point important avec les travaux de Kadanoff-Wilson, c'est que ***les lois de probabilité deviennent invariantes par changement d'échelle lors du changement de phase*** [22] (ex. liquide/vapeur, ferro/para). Ce point va nous intéresser tout particulièrement pour comprendre les réseaux de neurones.

Maintenant malgré tous ces très beaux développements, la compréhension d'un phénomène comme la turbulence qui a en apparence est simple reste à un stade embryonnaire

19. Voir https://www.insmi.cnrs.fr/fr/cnrsinfo/les-travaux-dhugo-duminil-copin

20. Il est aussi connu en Physique des Particules pour avoir établi en 1977 avec Guido Altarelli les équations éponymes d'évolutions des densités de partons avec l'échelle d'énergie en QCD, par la suite appelées équations DGLAP reconnaissant la valeur de travaux antérieur de Dokshitzer, Gribov et Lipatov.

21. Le mécanisme de Brout–Englert–Higgs–Guralnik–Hagen–Kibble plus précisément du nom de Peter Ware Higgs (1929-) associé à Robert Brout (1928-2011) François, Baron Englert (1932-) Gerald Guralnik (1936-2014), Carl Richard Hagen (1937-), et Thomas Walter Bannerman Kibble (1932-2016). Higgs et Englert recoivent le Noble de Physique en 2013 après la découverte de la particule H^0 au CERN par les équipes des expériences Atlas et CMS.

22. nb. les autres points essentiels sont la notion d'ordre et la notion de symétrie.

malgré la théorie de Kolmogorov, alors que d'un autre coté avec les réseaux de neurones on s'attaque à des problèmes bien plus complexes comme par exemple la génération de visages. Donc, on a l'impression que la Physique Statistique est restée à des phénomènes beaucoup plus simples, cependant quand on fait un pas, on a compris le pourquoi du comment, alors que l'on a l'impression si l'on peut dire qu'en Machine Learning on brûlerait les étapes.

2.6 Point de vue classique de la modélisation

Le sujet de la modélisation des données est un domaine à part entière qui fait partie des Prob. Stat., et ce faisant beaucoup de modèles ont été proposés. La première idée qui vient à l'esprit quand on aborde une image, un son (signal x en général) en constatant qu'il y a manifestement de la ***structure***, est que le signal x n'est pas un point arbitraire de $\mathbb{R}^d$ (d le nombre de variables, ex. nombre de pixels). Ainsi, le nombre de degrés de liberté m est *a priori* bien plus petit que d (ie. $m \ll d$), et on va modéliser alors ***un signal comme un élément d'une surface/variété*** $S \subset \mathbb{R}^d$ (Fig. 3). Ces modèles sont de ***type déterministe*** où la géométrie peut être illustrée en donnant une structure riemannienne à la variété, et en étudiant les plans tangents qui définissent ainsi une carte locale paramétrée avec des variables s, telles que le signal x est repéré selon $x = g(s)$. Ce qu'il faut dire tout net est que ***ce point de vue ne fonctionne absolument pas en grande dimension***. Premièrement, même si $m \ll d$, il n'en reste pas moins que m est grand [23], par exemple si l'on procède à une compression d'une image de 10^6 pixels, on arrive à un facteur 10 de réduction sans trop dégrader l'image, mais ce n'est qu'une réduction minime de la dimensionalité. Or si, m est la dimension de la variété, par un argument déjà développé dans les cours de 2018-19, afin d'estimer la géométrie de la variété, il faut procéder à un échantillonnage de celle-ci, et dans le cas simple d'un échantillonnage régulier de $[0, 1]^m$ avec une distance de $1/10$ entre les points, il nous faut 10^m points. Donc avec $m = 10^5$, on dépasse allègrement le nombre d'atomes dans l'Univers estimé à 10^{82}, autrement dit on se trouve face à ***la malédiction de la grande dimension***. On ne peut donc estimer avec précision la géométrie de la variété, c'est-à-dire que dans ce cadre ***l'on ne peut utiliser tous les outils de la géométrie différentielle***.

23. La "grande" dimension peut commencer dès $m = 10^3$.

Un autre point de vue consiste à prendre une ***approche probabiliste*** pour laquelle x devient la réalisation d'un vecteur aléatoire qui va avoir une certaine distribution $d\mu(x)$ que l'on supposera régulière par rapport à la mesure de Lebesgue, donc elle admettra une ***densité de probabilité*** $p(x)$ telle que $p(x)dx = d\mu(x)$, et le jeu va consister à trouver $p(x)$ qui décrit les données. Avec la ***Théorie de l'Information***[24], ce type de modélisation est beaucoup plus riche en grande dimension, et d'une certaine manière on peut se représenter la variété S comme une forme de ***géométrie floue.*** Les *théorèmes de concentration* donnent accès aux régions où les données ont une probabilité 1 de s'y trouver, mais ce faisant nous disposerons d'outils mathématiques beaucoup plus flexibles.

Si on revient un instant à la Physique Statistique en se demandant pourquoi il y a eu beaucoup de résistance aux idées de Boltzmann et Gibbs. En fait, il faut se rendre compte qu'il pouvait paraître choquant à ceux baignant dans le formalisme hamiltonien, que des ***lois déterministes étaient le résultat de processus stochastiques décrits par le formalisme de la Mécanique Statistique*** (même sans aller du coté du quantique). Bien entendu, le résultat de la loi des grands nombres de corpuscules fournit la solution pour retrouver les lois interprétées comme déterministes. Or ce type démarche va s'opérer dans le fait que des lois de probabilités $p(x)$ non uniformes, vont le devenir quand la dimensionalité d du problème va tendre vers l'infini. ***Ainsi par analogie avec la limite thermodynamique, on va retrouver des phénomènes de concentration autour de variétés de probabilités uniformes***.

2.7 Exemple des gaussiennes

2.7.1 Le cas où les variables ont même variance

Nous allons voir à travers cet exemple de distribution gaussienne un panorama de la problématique que nous approfondirons dans le cas non-gaussien. Ceci dit, le cas gaussien n'est pas juste un exemple d'école, il est très important à cause du théorème central limite. Donc, mettons que x a d variables $x = \{x_i\}_{i\leq d}$, et que les variables x_i sont ***indépendantes***

24. NDJE. Le lecteur pourra se rapprocher du cours de 2022.

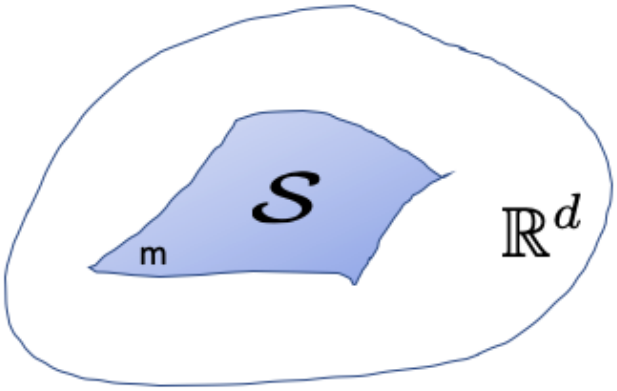


FIGURE 3 – Une image, un son, etc ayant une structure, est possiblement un élément d'une surface de $\mathbb{R}^d$ dont le nombre de degrés de liberté est bien plus petit.

identiquement distribuées (*iid* par la suite) de loi gaussienne $\mathcal{N}(\mu, \sigma^2)$. Ainsi,

$$p(x) = \prod_{i \leq d} p(x_i) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2}\sum_i \left(\frac{x_i - \mu}{\sigma}\right)^2\right) \tag{3}$$

La probabilité est maximale quand toutes les composantes x_i sont égales à μ. Maintenant où se trouve une réalisation de x? Le résultat n'est pas intuitif mais, x va se trouver sur une quasi-sphère définie par

$$\|x - \mu\|^2 = d\sigma^2 \tag{4}$$

"quasi" car l'épaisseur ε de la sphère est très petite comme on va le voir. Au passage, l'argument de l'exponentielle est appelé l'*énergie*[25]:

$$U(x) = \frac{1}{2}\sum_i \left(\frac{x_i - \mu}{\sigma}\right)^2 \tag{5}$$

25. nb. d'un point de vue notation, $E(x)$ aurait pu se confondre avec l'espérance que nous utiliserons abondamment, mais d'un autre coté U est souvent employé en physique pour l'énergie interne.

elle est toujours positive, et vaut 0 pour l'état le plus probable. La constante de proportionalité est notée Z^{-1} de telle sorte que

$$p(x) = Z^{-1} e^{-U(x)} \tag{6}$$

Considérant la somme suivante

$$S_d(x) = \frac{1}{d} \sum_{i=1}^{d} \left(\frac{x_i - \mu}{\sigma} \right)^2 \tag{7}$$

comme on a affaire à une somme de variables indépendantes [26]

$$\mathbb{E}_{x \sim p(x)}[S_d] = \mathbb{E}_{z \sim \mathcal{N}(\mu, \sigma^2)} \left[\left(\frac{z - \mu}{\sigma} \right)^2 \right] = 1 \tag{8}$$

tandis que

$$\mathbb{V}[S_d] = \frac{1}{d^2}(d \times 2) = \frac{2}{d} \tag{9}$$

donc la variance tend vers 0 pour les grandes dimensions. ***On a bien un phénomène de concentration, qui fait que le modèle probabiliste est très proche d'un modèle déterministe car la surface est quasi d'épaisseur nulle, mais pas négligeable et c'est exactement un exemple d'ensemble micro-canonique en Physique Statistique***.

2.7.2 Le cas général gaussien avec covariance

Dans ce modèle, la probabilité [27] $p(x)$ est définie selon l'expression suivante

$$p(x) = \mathcal{N}(\mu, C) = Z^{-1} \exp\Big(-\frac{1}{2}(x - \mu)^T C^{-1} (x - \mu) \Big), \qquad Z = (2\pi)^{d/2} |\det(C)|^{1/2} \tag{10}$$

où C est la matrice de covariance des variables $\{x_i\}_{i \leq d}$. C est diagonalisable avec des valeurs propres positives $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ et il a apparait des variables "réduites" $\{z_i\}_{i \leq d}$ qui sont les axes principaux d'un ellipsoïde à d dimension. Ainsi le passage de x à z permet

26. $((z - \mu)/\sigma)^2$ avec $z \sim \mathcal{N}(\mu, \sigma^2)$ suit une loi du χ^2 à 1 degré de liberté, dont la moyenne est 1 et la variance 2.

27. On omet généralement le terme "densité"...

de réécrire $p(x)$ selon

$$p(z) = Z^{-1} \exp\left(-\frac{1}{2}\sum_i \left(\frac{z_i}{\sigma_i}\right)^2\right), \qquad Z = (2\pi)^{d/2}\prod_i \sigma_i \tag{11}$$

Il n'est pas très difficile de généraliser le cas précédent (cf. les d variables $(z_i/\sigma_i)^2$ avec $z_i \sim \mathcal{N}(0, \sigma_i^2)$ suivent une loi du χ^2 à 1 degré de liberté), et l'on se rend compte que les données vont donc se concentrer sur des ***coquilles ellipsoïdales*** centrées sur μ et dont les extensions le long des axes principaux sont données par les valeurs des σ_i et dont l'épaisseur va décroitre en $1/d$.

Maintenant, ce modèle à covariance permet ***de sélectionner les composantes les plus pertinentes*** en ne prenant que celles ayant une valeur de σ_i au dessus d'un certain seuil ε (cf. analyse en composante principale, PCA). Ainsi, on peut faire de la ***réduction de dimension***, et de même on va trouver que ***la probabilité de se trouver sur ces coquilles ellipsoïdales va être uniforme***.

Les observations précédentes de concentration et d'uniformité de probabilité sur des ensembles "typiques" sont à la base de la***Théorie de l'Information***, et la raison pour laquelle l'***entropie*** joue un rôle si fondamental pour comprendre ces ***phénomènes de grande dimension***.

2.8 Le point de vue de la Théorie de l'Information hors cas gaussien

Si l'on généralise à un cas non gaussien, tout en conservant l'hypothèse fondamentale d'*iid*, l'on a une convergence en probabilité (loi faible des grands nombres) telle que

$$-\frac{1}{d}\log p(x_1, x_2, \ldots, x_d) = -\frac{1}{d}\sum_i \log p(x_i) \xrightarrow[d\to\infty]{} \mathbb{E}_{x\sim p(x)}[-\log p(x)] = \bar{\mathbb{H}} \tag{12}$$

où l'on note [28] $\bar{\mathbb{H}}$ la densité d'entropie associée à la distribution $p(x)$ (ici d'une des d composantes, c'est-à-dire que $d \times \bar{\mathbb{H}}$ est l'entropie totale associée au système à d composantes *iid*, d jouant le rôle identique au nombre N de corpuscules d'un système thermodynamique,

28. NDJE. j'essaye de conserver des notations cohérentes avec le cours de 2022 mais il se peut qu'il y ait des variations.

en d'autres termes l'entropie est une variable extensive). Ainsi,

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} \mathbb{P}\left(\left|-\frac{1}{d}\log p(x_1, x_2, \ldots, x_d) - \bar{\mathbb{H}}\right| \leq \varepsilon\right) = 1 \tag{13}$$

On rejoint l'idée de la surface S définit par une équation implicite $f(x) = 0$ où l'on prendrait

$$f(x) = \left|-\frac{1}{d}\log p(x_1, x_2, \ldots, x_d) - \bar{\mathbb{H}}\right| \tag{14}$$

Il n'est pas exact strictement que $f(x)$ soit nulle, car il y a une épaisseur ε, mais on retrouve l'idée de concentration de $-\log p$ vers l'entropie qui va entrainer une concentration des données sur une surface, et l'on verra que la probabilité est uniforme sur celle-ci. C'est le point de vue développé [29] en 1948 par Claude Shannon (1916-2001) qui donne lieu à la notion d'***ensembles typiques*** que nous retrouverons cette année.

Ainsi, les notions suivantes se dégagent:

- l'énergie $E(x)$ qui est très similaire (à une constante près $\log Z$) à $-\log p(x)$;
- l'entropie $\mathbb{H}$ qui n'est autre que l'espérance $\mathbb{E}[-\log p(x)]$, laquelle va nous permettre d'appréhender la complexité du système (cf. le nombre de configuration à la manière de Boltzmann).

Finalement, ces outils de la Physique Statistique et de la Théorie de l'Information nous permettent de penser la grande dimension. Mais ***la grande difficulté est de construire*** $p(x)$ c'est-à-dire de concevoir des modèles qui s'appliquent aux données lesquelles n'ont rien de gaussien la plupart du temps. Dans le cas de la Physique (gaz parfait/Ising) on a des modèles des interactions assez locales entre corpuscules/spins, ce qui n'est absolument pas le cas en ***Science des Données où l'on n'a pas de modèle***. En effet, quel peut bien être le modèle sous-jacent des images de chien/chat? Telle est *la* question si l'on peut dire.

2.9 Les modèles non gaussiens stationnaires ergodiques

Dans la suite on va se concentrer sur ***des modèles stationnaires et ergodiques*** [30]. L'ergodicité stipule que la valeur moyenne d'une grandeur calculée de manière statistique (mathématiquement accessible) est égale à la moyenne d'un très grand nombre de mesures

29. Cours 2022
30. Hypothèse formulée par Boltzmann en 1871

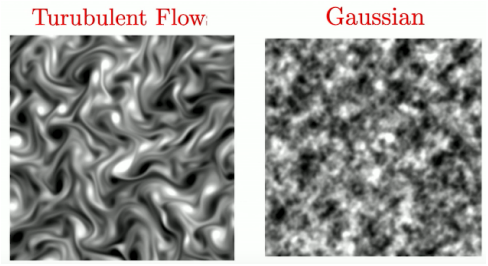


FIGURE 4 – Exemple d'une simulation d'un écoulement turbulent en 2D (gauche), et de la réalisation (droite) d'un modèle gaussien qui simplifie le problème en ne prenant en compte que les moments d'ordre 2 de la distribution de gauche c'est-à-dire en ne considérant que la matrice de covariance.

prises dans le temps (accessible par l'expérience). Dans le ***cas gaussien***, de ces deux hypothèses découle, comme on va le voir, que des phénomènes qui se déroulent à deux ***échelles distinctes sont indépendants***. Par contre, dans le cas ***non gaussien***, on voit apparaitre de la structuration (texture, tourbillon) du fait que des phénomènes se déroulant à deux ***échelles distinctes ne sont pas indépendants***. C'est bien cette ***notion de dépendance*** qu'il va nous falloir comprendre afin de construire des modèles non gaussiens.

2.9.1 La turbulence et modèle Ising

Pour comprendre la non gaussianité, S. Mallat donne quelques images que je me permets de reproduire ici. Par exemple, sur la figure 4 on distingue sur la vue de gauche les vortex d'un fluide turbulent. Le modèle gaussien (vue de droite) ne prenant en compte que les moments d'ordre 2 (ie. la matrice de covariance) non seulement ne reproduit pas cette structuration, mais aussi est beaucoup plus désordonné. Ce désordre est le reflet d'une plus grande entropie. Donc, l'enjeu de la Physique Statistique a été de trouver des modèles non gaussiens pour rendre compte de cette structuration. Cependant, il faut reconnaitre que cela n'a pas été couronné de succès jusqu'à encore très récemment.

Un modèle numérique qui illustre les propriétés de transition de phase ferro/paramagnétisme est celui d'Ising. Brièvement, le système est constitué de N *spins* σ à deux composantes $\{-1, +1\}$ placés aux nœuds d'un réseau (ex. à base carré). Si deux spins sont voisins, ils subissent une interaction de constante de couplage J et pour simplifier, on

définit une matrice symétrique d'interaction J_{ij} qui n'a que des éléments non nuls (égaux à J que l'on prend positif ici) pour uniquement un couple (i, j) de plus proches voisins. De plus, on plonge le système dans un champ H externe que l'on suppose homogène. Ainsi, chaque spin beigne dans un champ magnétique induit non seulement par H mais aussi par l'action collectives des autres spins du réseaux.

L'énergie du système est alors donnée par l'expression [31]

$$U(\boldsymbol{\sigma}, J) = -\frac{1}{2} \sum_{(i,j)ppv} J_{ij}\sigma_i\sigma_j - h\sum_i \sigma_i \qquad \boldsymbol{\sigma} \in \{-1, +1\}^N \tag{15}$$

L'on définit alors l'aimantation moyenne d'une configuration "c" de spins par

$$S_c(N, h) = \frac{1}{N}\sum_{i=1}^{N} \sigma_i \tag{16}$$

Remarquons que si $h = 0$, il y a une *symétrie du système* se traduisant par le fait que l'énergie se conserve par renversement de touts les spins. Le système à l'équilibre se met dans une configuration d'énergie minimale avec soit tous les spins orientés $+1$, ou bien tous orientés -1. L'aimantation moyenne est égale soit à $+1$ soit -1 ($S_c(N, h = 0) = \pm 1$). Si h est très intense, alors l'énergie minimale est gouvernée par le second terme et la configuration des spins est obtenue en orientant tous les spins dans le même sens que le champ local, dans ce cas il y a une aimantation forcée du système. Le champ externe fixe une direction et l'on peut la choisir comme axe pour évaluer les spins, donc $h > 0$ par convention. On a brisé la symétrie, et $S_c(N, h \gg 1) = +1$.

Maintenant, mettons le système en contact avec un réservoir qui fixe la température T. Intuitivement, on se dit qu'à haute température l'agitation thermique va rendre aléatoire l'orientation des spins même avec $h \neq 0$, et il n'y a pas d'aimantation qui persiste. Il y a une compétition entre température (*désordre*) et champ externe (*ordre*). Mais que dire de la quantité:

$$\lim_{h\to 0}\lim_{N\to\infty} S_c(h, N) = M_0 \tag{17}$$

appelée *aimantation spontanée*? C'est-à-dire ***que se passe-t-il à la limite thermodyna-***

31. Notons que si H est inhomogène, la famille de modèles d'Ising est connue sous le nom de *Hopfield networks* ou *Boltzmann machines* dans la communauté des réseaux de neurones.

mique quand on coupe le champ externe[32]? La réponse dépend de la température: si elle est inférieure à une ***température critique*** $T < T_c$ alors il subsiste une aimantation spontanée non nulle (Fig. 5). Qui plus est l'on constate qu'il y a de grandes plages (régions) où les spins ont la même orientation. A température nulle, on se retrouve avec la configuration où les spins sont tous alignés dans la même direction.

C'est toute l'étude des transitions de phase que l'on peut alors étudier avec ce modèle emblématique. En particulier, ***comment des corrélations à longue portée émergent dans le réseau alors que l'on a des interactions uniquement entre plus proche voisins*** (interactions *locales*)? Une fois ce problème compris par la Physique Statistique (>1950), on a essayé de l'appliquer dans le cas de la turbulence. Malheureusement de nouveau c'est un échec. C'est-à-dire qu'en essayant de construire des interactions locales, on ne parvient pas à comprendre l'apparition des structures de la figure 4.

2.9.2 La génération de trames sonores

NDJE. S.Mallat fait écouter des trames de son, donc je vous engage à les (ré)écouter dans la vidéo du cours vers 1:14:00 du début. Egalement pour comprendre les notions de représentations Temps-Fréquence, je vous engage vous reporter au cours de 2020 sur la Transformée de Fourier (notamment à fenêtre) et la Transformée en Ondelettes dont l'algorithme de décomposition est détaillé aussi dans le cours de 2021.

Nous avons sur la figure 6 la représentation Temps-Fréquence de quatre séquences de musique: note continue, "attaque", "trémolo", "vibrato". On voit distinctement les différences de ces quatre trames non seulement à la fréquence fondamentale mais aussi dans les harmoniques. Une fois que l'on a saisi comment se construit de telles représentations, la figure 7 présente 3 représentations d'un "applaudissement": le son original (vue de gauche) présente des structurations à la fois fréquentielles et temporelles; puis sur la vue du milieu une génération par *un modèle gaussien* qui ne prend en compte que la matrice de covariance le long du temps (l'énergie de chaque harmonique est respectée), il n'y a *aucun phénomène transitoire ni de corrélations entre échelle* comme on peut les voir sur le son original; enfin sur la vue de droite la génération par un modèle qui prend en compte les corrélations entre échelles, le modèle certes pas parfait est néanmoins beaucoup plus

32. nb. je vous laisse deviner la réponse si on procède à l'inversion des deux limites.

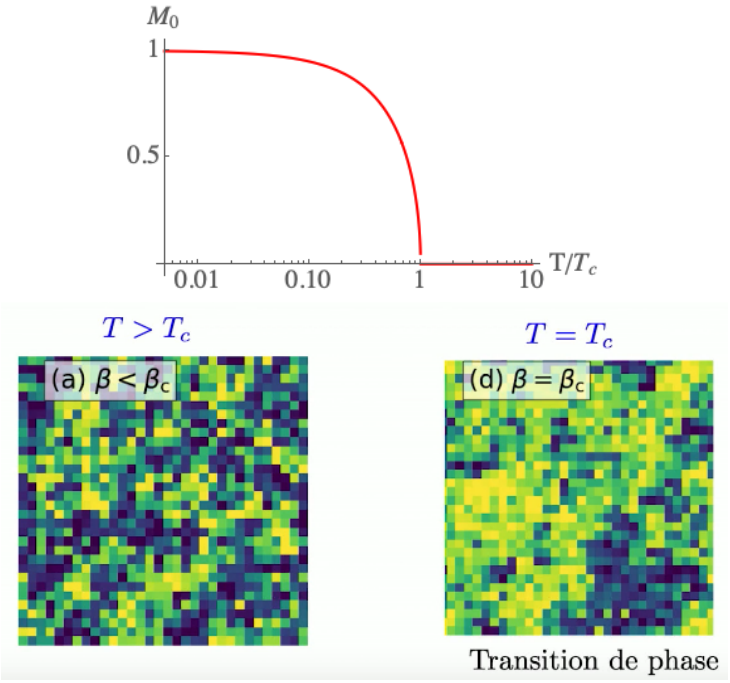


FIGURE 5 – (haut): Aimantation spontanée M_0 dans le cas d'un modèle d'Ising en fonction de la température. Exemple d'une configuration simulée d'un réseau de spins (Ising ici où les spins prennent des valeurs continues) soit mis au contact d'un réservoir dont la température est au-dessus de la température critique (vue de gauche) ou en deçà (vue de droite). On peut donc apprécier le changement de structuration qui s'opère lors de la transition de phase.

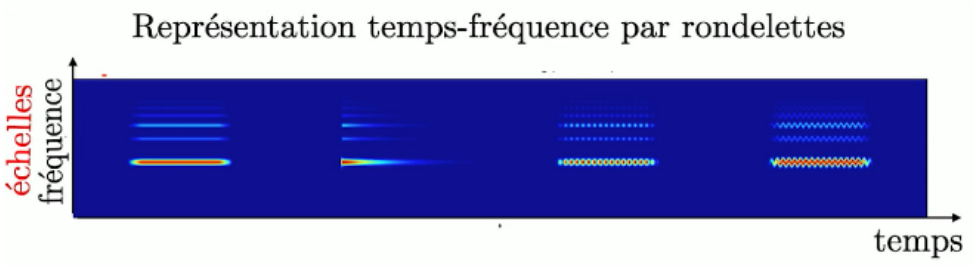


FIGURE 6 – Représentation Temps-Fréquence de quatre séquences de musique.

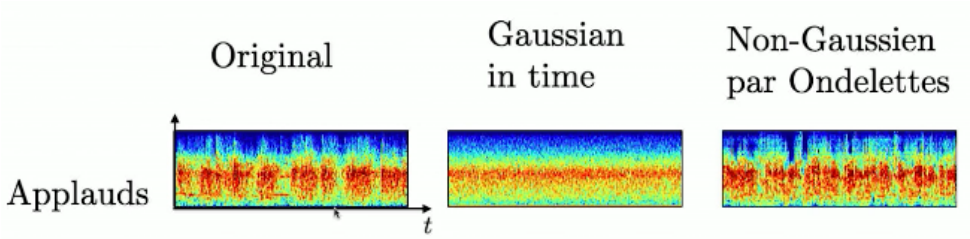


FIGURE 7 – Représentations Temps-Fréquence d'un applaudissement: le son original à gauche; au milieu la génération d'un modèle gaussien qui ne conserve que l'énergie originale à toutes les échelles/fréquences; à droite la génération par un modèle plus réaliste qui tient compte des corrélations entre échelles.

réaliste. Nous verrons comment construire de tels modèles et en quoi par exemple les non-linéarités (ex. ReLU) sont fondamentales.

2.9.3 Autres exemples

Ensuite une fois que l'on a compris comment modéliser des interactions non-locales avec les corrélations entre échelles, on peut par exemple générer des images de champ non-gaussiens comme illustré sur la figure 8. On peut également faire de la séparation de composantes par exemple en séparant dans une image constituée à partir de photons collectés par le satellite Planck venant de la voute céleste, la partie gaussienne issue du fond diffus cosmologique, et la partie non gaussienne constituée par ce que l'on nomme des avants-plans (ex. émission par des poussières intergalactiques) (Fig. 9). Ceci fera l'objet du séminaire d'E. Allys.

D'un autre coté, nous n'aborderons pas cette année, nous dit S. Mallat, les modèles non-stationnaires et non-ergodiques comme la génération d'images de visages ou de phrases parlées. Les réseaux de neurones savent faire mais les mathématiques sont plus compliquées.

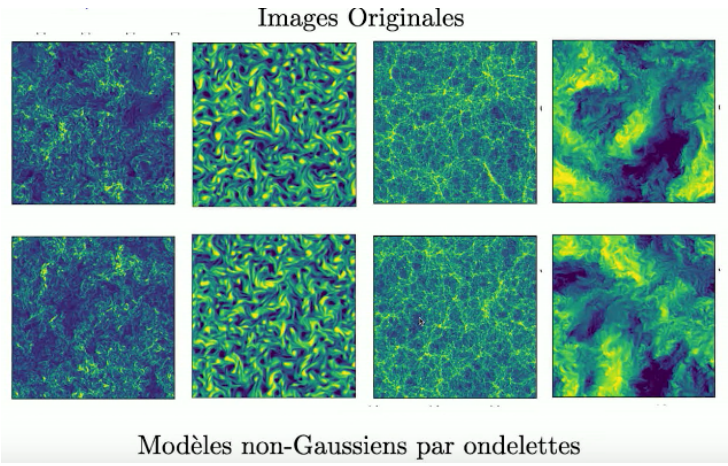


FIGURE 8 – Exemple de génération de champs non gaussiens par des modèles par ondelettes: turbulence, *cosmic web*, etc.

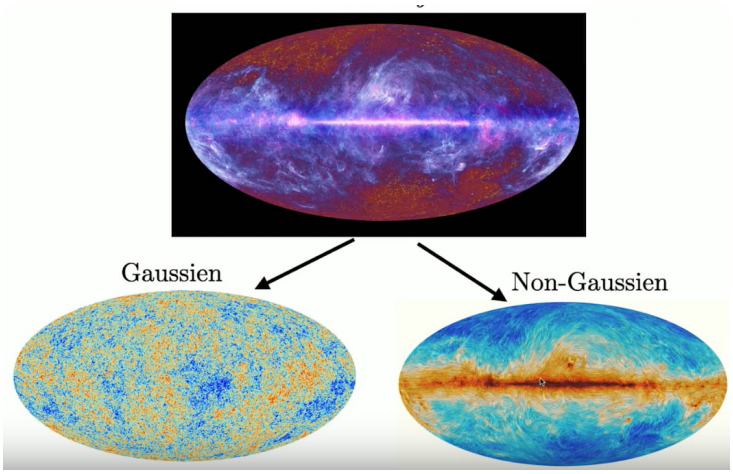


FIGURE 9 – Exemple de séparation de composantes gaussiennes/non-gaussiennes appliquée à la cosmologie.

2.10 Plan du cours [33] 2023

- Nous commencerons par une ***brève introduction à la modélisation***, avec un retour sur la malédiction de la dimension, et l'introduction aux champs de Markov et modèles à entropie maximale.
- Ensuite, nous passerons en revue le point de vue de la ***Physique Statistique***: les *modèles Micro-canoniques* où nous verrons le lien entre *énergie*, *entropie* et *température*, ainsi que la notion d'*équilibre thermodynamique*.
- De ces idées de Phys. Stat., nous allons envisager le point de vue de la ***Théorie de l'Information*** afin de mettre en évidence les correspondances. Ce faisant, nous revisiterons la notion d'entropie en lien avec les *probabilités*, et détaillerons les notions d'*entropie différentielle* et d'*entropie conditionnelle*/relative et "distance" de *Kullback-Leibler*, ainsi que la *densité d'entropie*.
- Pour comprendre la notion d'équilibre, nous allons utiliser les ***chaines de Markov*** qui vont nous donner accès à 2 choses: premièrement on va appréhender l'entropie dans le cas où il y a une *forme de dépendance*, et deuxièmement comment se passe l'*évolution vers l'équilibre* avec par exemple le *2nd Principe de la Thermodynamique* que nous démontrerons sous l'hypothèse markovienne. Nous verrons également comment on peut échantillonner les distributions avec ***les algorithmes Monte Carlo*** avec les chaines de Markov (MCMC) très importants en pratique.
- Nous reviendrons alors à la notion de modèle, par le biais de ***Modèles à Entropie Maximum***. Ainsi nous verrons entre autres la notion de *dualité* sur des problèmes convexes (transformation de Legendre), *les modèles Macro-canoniques*, etc.
- Finalement, nous développerons des ***modèles non gaussiens*** avec toute l'***Analyse Harmonique***. Nous visiterons la *Transformée de Fourier* (processus stationnaires qui nous permettent de diagonaliser la covariance), la *Transformée en Ondelettes* (Analyse multi-échelles), et toutes les *interactions non-linéaires*.

33. NDJE. Gardez sous le coude les notes des cours de 2020, 2021 et 2022.

3. Séance du 25 Janv.

Durant cette séance nous allons en premier lieu nous poser la question de *comment estimer un modèle en grande dimension?* Nous verrons alors les stratégies pour contourner *la malédiction de la dimension*. En second lieu, nous verrons le même problème dans le cadre de la *Physique Statistique*. Rappelons que l'enjeu est d'estimer une probabilité $p(x)$ où $x \in \mathbb{R}^d$ avec $d \gg 1$ (typiquement $d \sim 10^5 - 10^6$).

Pour fixer le cadre de l'exemple que nous allons étudier, on fixe $x \in \{-1, +1\}^d$ qui peut par exemple modéliser le résultat d'un questionnaire binaire (Oui/Non) en vue d'un diagnostic médical, ou bien la valeur de d spins d'un modèle d'Ising.

L'ensemble $\mathcal{F}$ des probabilités $p(x)$ est tel que

$$\mathcal{F} = \left\{ p : \{-1, +1\}^d = \mathcal{E}_d \to [0, 1], tq. \sum_{x_i \in \mathcal{E}_d} p(x_i) = 1 \right\} \tag{18}$$

On peut alors se poser les problèmes suivants:

- Il va nous falloir restreindre la recherche à des sous classes de $\mathcal{F}$, car sans information *a priori* nous allons être confronté à des difficultés. Ainsi, il va nous falloir faire de la ***modélisation***, laquelle est souvent paramétrée.
- une fois définie la sous classe $\mathcal{C} \subset \mathcal{F}$, nous allons aborder le problème d'***estimation*** des paramètres du modèle, c'est-à-dire le problème d'***apprentissage***.
- Nous verrons les problèmes d'***inférence***, comme par exemple le calcul des *probabilités marginales* (ex. quel est le diagnostic pour telle ou telle maladie) qui nécessite alors des intégrations en grande dimension.
- Finalement, on peut vouloir ***générer*** de nouvelle données typiques une fois connue $p(x)$, c'est un problème d'***échantillonnage***.

3.1 Approche fréquentiste

Mettons que l'on dispose d'échantillons, on peut estimer les probabilités en utilisant des histogrammes. Soit $\mathcal{E}_d$ l'ensemble des possibilités d'un système à d variables, dont

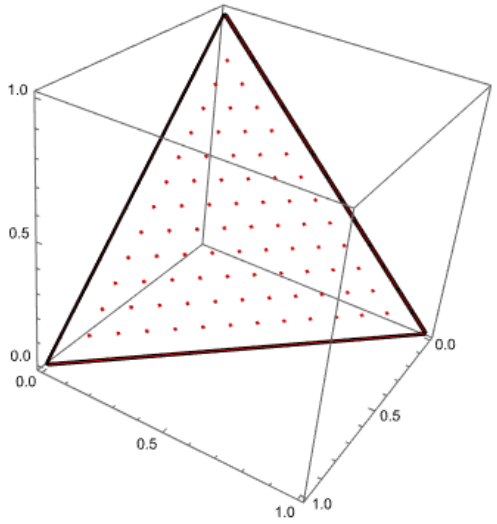


FIGURE 10 – Lieu des solutions $\sum_k \theta_k = 1$ (ici dans le cas de 3 paramètres).

chacune est binaire $y \in \{-1, +1\}$, alors

$$\mathcal{E}_d = \{y_1, y_2, \ldots, y_K\}, \qquad K = 2^d \tag{19}$$

Et l'on veut estimer $p(x = y_k)$ que l'on note θ_k, ainsi on a K paramètres. Ce que l'on sait c'est que

$$\sum_{k=1}^{K} \theta_k = 1 \tag{20}$$

D'un point de vue géométrique le K-uplet recherché se trouve sur le "simplex" (Fig. 10). Nous avons à notre disposition n données/échantillons [34] $\{x^i\}_{i \leq n}$. Par hypothèse, les données sont *iid* selon la densité $p(x)$ recherchée, se sont des exemples d'un vecteur aléatoire. L'estimateur $\hat{\theta}_k$ de θ_k est donnée par [35]

$$\hat{\theta}_k = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{x^i = y_k\}} \tag{21}$$

34. nb. l'exposant i indique la i-éme donnée, sachant que c'est un vecteur à d composantes.
35. $\mathbf{1}_A$ est l'indicatrice de l'ensemble A.

Il s'agit du k-ième bin de l'histogramme, aussi appelée la distribution empirique. Comme les x^i sont des *v.a* (*variable aléatoire*), on peut étudier la *v.a* suivante

$$z_k = \mathbf{1}_{\{x^i = y_k\}} \tag{22}$$

qui peut prendre les valeurs 0 ou 1, c'est-à-dire que chaque z_k est une variable de Bernoulli. Ainsi, l'espérance de z_k est donnée par

$$\mathbb{E}[z_k] = \theta_k \tag{23}$$

(en effet, c'est l'espérance des "1") et donc sa variance est

$$\mathbb{V}[z_k] = \theta_k(1 - \theta_k) \tag{24}$$

Maintenant, concernant l'estimateur $\hat{\theta}_k$ on peut également calculer sa moyenne et sa variance, il vient:

$$\mathbb{E}[\hat{\theta}_k] = \theta_k \qquad\qquad \mathbb{V}[\hat{\theta}_k] = \mathbb{E}[|\hat{\theta}_k - \theta_k|^2] = \frac{\theta_k(1-\theta_k)}{n} \tag{25}$$

il est non-biaisé, et la variance décroit bien avec n. Mais, ce qui nous intéresse c'est de savoir si on estime bien la probabilité $p(x)$. Alors, si l'on note $\Theta = (\theta_k)_{k \leq K}$ (idem $\hat{\Theta}$), il nous faut estimer les propriétés du vecteur normalisé suivant

$$\frac{\mathbb{E}[\|\Theta - \hat{\Theta}\|^2]}{\|\Theta\|^2} = \frac{\sum_k \mathbb{E}[|\theta_k - \hat{\theta}_k|^2]}{\|\Theta\|^2} = \frac{\sum_k \theta_k(1-\theta_k)}{n\|\Theta\|^2} = \frac{1 - \|\Theta\|^2}{n\|\Theta\|^2} \tag{26}$$

Donc la question maintenant est de savoir si on prend au hasard un Θ sur le simplex (Fig. 10) quelle est l'erreur moyenne? En grande dimension, la localisation moyenne de Θ est au centre du symplex c'est-à-dire $\forall k, \theta_k = 1/K$ ce qui donne $\|\Theta\|^2 = K/K^2 = 1/K$, d'où

$$\frac{\mathbb{E}[\|\Theta - \hat{\Theta}\|^2]}{\|\Theta\|^2} \approx \frac{K}{n} = \frac{\text{nbre de paramètres}}{\text{nbre d}'\text{exemples}} \tag{27}$$

Or si l'on veut une bonne estimation cela signifie que le nombre d'exemples doit être grand selon la loi d'échelle

$$n \gg 2^d \tag{28}$$

que dire alors dans le cas où $d = 10^6$! Et même avec $d = 100$, il est clair que l'on est confronté à un problème de ***malédiction de la dimensionalité***. Conclusion: ***l'approche fréquentiste ne marche pas en grande dimension***.

Il faut réaliser cet état de fait car il est d'usage courent d'utiliser cette approche en histogrammant les données pour obtenir des "probas expérimentales", donc dans le cas de la grande dimension il faut aller à l'encontre de cette "intuition". Fort heureusement, il y a un autre point de vue, à savoir l'***approche bayésienne***[36] qui considère les probas sous l'angle de mesure d'incertitude ou de représentation d'information partielle.

3.2 Approche bayésienne, modèle Markov local

C'est le point de vue dominant en grande dimension. Nous allons bâtir le modèle à partir des ***dépendances connues entre les variables*** et l'on essaye de représenter l'*incertitude sur le processus sous-jacent* qui a délivré les données dont on dispose.

Ceci étant, si l'on a deux *v.a*, la probabilité jointe peut se décomposer de la sorte selon l'idée de Bayes

$$p(x_1, x_2) = p(x_1)p(x_2|x_1) \tag{29}$$

ce qui se généralise pour d variables selon

$$\begin{aligned} p(x_1, \ldots, x_d) &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\ldots p(x_d|x_1, \ldots, x_{d-1}) \\ &= p(x_1)\prod_{t=2}^{d} p(x_t|x_1, \ldots, x_{t-1}) \end{aligned} \tag{30}$$

Pour le moment, on ne voit pas de gain à faire cette décomposition. Là où l'on gagne c'est si l'on peut éliminer des variables dans l'ensemble $(x_1, \ldots, x_{t-1})$ pour conditionner x_t. C'est la notion d'***indépendance conditionnelle*** qui est fondamentale dans le processus de ***la réduction de dimensionalité***, afin de casser la malédiction citée à la section précédente. Un point de notation:

$$p(x_3|x_1, x_2) = p(x_3|x_2) \Leftrightarrow x_2 \perp x_3|x_1 \tag{31}$$

36. NDJE. voir également la Sec. 2.2 du cours de 2022.

Si l'on peut identifier de telles sources d'indépendance permettant de telles réductions en enlevant suffisamment de variables, alors chaque probabilité conditionnelle est un problème en basse dimension plus accessible.

Examinons un exemple "extrême" de ce type d'indépendance à savoir ***les chaînes de Markov***. Dans ce cas, il n'y a pas de mémoire se traduisant par

$$p(x_t|x_1, \ldots, x_{t-1}) = p(x_t|x_{t-1}) \tag{32}$$

et donc

$$p(x) = p(x_1) \prod_{t=2}^{d} p(x_t|x_{t-1}) \tag{33}$$

Les probabilités conditionnelles (- de transition) sont des fonctions à deux variables. Si de plus, nous sommes en régime stationnaire, celles-ci sont toutes identiques. Donc, dans ce cas le problème est relativement aisé, mais il s'agit d'un cas très particulier.

D'une manière plus générale, les ***modèles graphiques*** sont des modèles où les dépendances conditionnelles entre variables sont organisées selon un *graphe*. C'est une forme de *topologie* donnée *a priori* sur les données. Un exemple est celui du modèle d'Ising où l'on ne considère que les interactions entre plus proches voisins. D'une manière un peu plus générale, les dépendances (conditionnnelles) seront cependant *locales*:

$$p(x_i|x_{j\neq i}) = p(x_i|x_j \ tq. \ j \in \mathcal{C}_i) \tag{34}$$

où $\mathcal{C}_i$ est appelée une "clique" (voisinage de i). Le nombre de variables qui vont finalement intervenir dans le produit des probabilités conditionnelles, va dépendre directement du nombre d'éléments dans les voisinages. On a notamment le théorème Hammersle-Clifford qui nous dit que

$$p(x) = \prod_{i=1}^{d} \psi(x_i, i \in \mathcal{C}_i) \tag{35}$$

Ce n'est qu'une autre façon d'écrire les probabilités conditionnelles (Eq. 30) en se restreignant aux voisinages à considérer. Si $|\mathcal{C}_i| \ll d$ et s'il y a une invariance dans le réseau, alors ***on peut gagner face à la malédiction de la dimension***.

Cependant, examinons le problème suivant: par exemple dans une image de la pièce qui vous entoure où l'on serait tenté de prendre $|\mathcal{C}_i| = 8$ (8 pixels entourant chaque pixel),

comment rendre compte de la structuration à grande échelle comme le fait qu'il y a des pixels très distants qui appartiennent à la frontières d'objets, et sont donc corrélés. Donc, ce type de modélisation marche uniquement dans certaines situations.

Par contre en physique/chimie, on peut ne considérer en première approximation que des interactions locales au niveau microscopique: ex. dans une molécule avec des liaisons chimiques covalentes σ, où des liaisons ioniques dans un cristal. Mais pour raffiner le modèle, il faut prendre en compte des interactions à longue portée: ex. forces dipôle-dipôle de Van der Waals en $1/r^7$, liaison chimique covalente π ou *hydrogène*.

Ce qui peut nous aider c'est la ***hiérarchisation*** dans l'intensité des différentes interactions, et dans ce contexte, le modèle des chaines de Markov peut présenter le modèle de base des interactions locales les plus intenses que l'on peut raffiner par l'adjonction des interactions plus faibles. Ce que l'on remarque dans ce cheminement, c'est que l'intuition de la physique (notion de ***multi-échelles***) parait très naturelle.

3.3 Modèles d'entropie maximum

Si dans la section précédente *la réduction de dimensionalité se fait en ne considérant que des structures locales* du système, nous allons procéder tout autrement en ne considérant que ***des mesures globales***. Cette stratégie est à la base de la Physique Statistique: on va décrire ***la distribution de probabilité en fonction de grandeurs macroscopiques***. Formellement, nous allons regarder des *moments* de la distribution de probabilité:

$$\mu_k = \mathbb{E}_{x\sim p(x)}[\phi_k(x)] = \int_{\mathbb{R}^d} \phi_k(x)p(x)dx \tag{36}$$

Remarquons au passage que l'intégration se fait en (très) grande dimension. Par exemple si $\phi_k(x) = x_k$ (k-ième coordonnée de x), $\mu_k = \mathbb{E}[x_k]$. Dans le cas *stationnaire* il y a invariance $\mu_k = \mu$, et en intégrant sur toutes les variables, on a une mesure globale. Autre exemple: $\phi_k(x) = x_i x_{i-k}$ (imaginez des pixels pris deux-à-deux mais dont tous les couples retenus sont formés de pixels distants de la même quantité k), dans le cas *stationnaire*, $\mathbb{E}[\phi_k(x)] = f(k)$, c'est-à-dire que la fonction f ne dépend pas de i. Bien entendu, on peut prendre des fonctions bien plus complexes. De ces moments, comment obtenir $p(x)$?

Bien entendu en général, il ne suffit pas de se donner un ou deux moments pour déterminer une distribution, mais on peut se demander ***quelle est la distribution la plus***

régulière qui redonne les moments dont on dispose? Dans ce contexte la question est de savoir quel est l'***indicateur de régularité*** qui va nous guider. La réponse est l'***entropie*** $\mathbb{H}[p]$ définie par

$$\mathbb{H}[p] = -\int_{\mathbb{R}^d} p(x)\log p(x)dx = \mathbb{E}[-\log p(x)] \tag{37}$$

Il nous faut donc trouver une distribution $p(x)$ qui satisfait des contraintes à la fois constituées par les moments $(\mu_k)_{k\leq K}$ et par la qualité de régularité/entropie maximum. Les paramètres du modèle de densité de probabilité p_θ sont les $(\mu_k)_{k\leq K}$, ainsi on cherche

$$p_\theta,\ tq.\ \forall k \mathbb{E}_{p_\theta}[\phi_k] = \mu_k,\ \text{et}\ \max_{p_\theta}(\mathbb{H}[p_\theta]) \tag{38}$$

Nous verrons plus en détails dans le cours que la solution à ce problème d'optimisation convexe peut s'écrire selon l'expression

$$p_\theta(x) = Z^{-1}\exp\left(-\sum_{k=1}^{K}\theta_k\phi_k(x)\right) \Leftrightarrow -\log p_\theta(x) = \log Z + \sum_{k=1}^{K}\theta_k\phi_k(x) \tag{39}$$

avec ***les*** θ_k ***les multiplicateurs de Lagrange***[37] associés aux contraintes. Z_θ est la constante de (re)normalisation[38] pour que l'intégrale de p_θ soit égale à 1 qui peut se mettre sous la forme ($\Theta = (\theta_k)_{k\leq K}$, $\Phi(x) = (\phi_k(x))_{k\leq K}$)[39]

$$Z_\theta = \int_{\mathbb{R}^d} e^{-\Theta^T\Phi(x)}dx \tag{40}$$

De cette modélisation, nous pouvons faire quelques remarques:

- nous avons deux façons de représenter la densité de probabilité: soit d'un coté les *moments* $(\mu_k)_{k\leq K}$, soit d'un autre coté les *paramètres* $(\theta_k)_{k\leq K}$;
- on peut voir le produit scalaire $\Theta^T\Phi(x)$ comme une *énergie*[40] $U(x)$. Or, en prenant l'espérance de l'expression de droite de Eq. 39, on obtient

$$\mathbb{H}[p_\theta] = \mathbb{E}_{x\sim p_\theta}[\log Z] + \mathbb{E}_{x\sim p_\theta}[U_\theta(x)] = \log Z + \bar{U}_\theta = -F_\theta + \bar{U}_\theta \tag{41}$$

37. NDJE. Voir aussi Sec. 8.3 Cours 2018.
38. nb. En Phys. Stat. c'est la fonction de partition de Gibbs.
39. nb. En ML le vecteur Φ est souvent appelé le *feature vector*.
40. nb. d'un point de vue notation, $E(x)$ aurait pu se confondre avec l'espérance, de plus U est employé en thermodynamique pour l'énergie interne.

où l'on à fait apparaitre, l'***entropie*** $\mathbb{H}[p_\theta]$, l'***énergie libre*** F_θ et l'***énergie interne*** moyenne $\bar{U}_\theta$ du système. C'est une relation que l'on va retrouver en Physique Statistique [41]. Cette relation peut s'écrire selon

$$\bar{U}_\theta = \mathbb{H}[p_\theta] + F_\theta \tag{42}$$

En thermodynamique, l'énergie libre $F = -\log Z$, un des outils introduits par Gibbs, dépend par exemple de la température, du nombre de corpuscules, du volume, de variables généralisés associée à des forces, et par dérivation elle permet de retrouver, l'énergie moyenne, la pression moyenne, la force moyenne qui s'exerce sur le système, etc. Ici nous voyons la vision statistique de ces concepts.

3.4 Inférence et applications des modèles probabilistes

3.4.1 Problèmes inverses

Une fois que l'on a un modèle probabiliste, on peut s'attaquer à tous les problèmes classiques de l'analyse de données. Le premier type de problème concerne des données/mesures x obtenues (1) à partir d'un signal y, transformé par exemple par un opérateur K que l'on suppose linéaire par simplicité du propos, et (2) bruitées (η une $v.a$). Ainsi l'on a la relation entre x et y suivante

$$x = Ky + \eta \tag{43}$$

Un exemple: supposons que l'on prenne des images avec un appareil photo muni d'un capteur CCD, chaque pixel constitué d'une photodiode silicium produit un signal électrique (valeur macroscopique) proportionnel à l'énergie déposée par les photons (au niveau microscopique) durant un cycle d'horloge. L'objectif alors est de récupérer y à partir de x pour obtenir par exemple des images de plus haute résolution à partir des moyennes locales. Un autre exemple en imagerie médicale: en utilisant les rayons X lors de radiographies prises sous différents angles, on obtient des projections de l'interaction matière-rayonnement

41. NDJE. l'équation en thermodynamique fait intervenir la température qui est paramètre, selon $TS = \bar{U} - F$.

(x), dans ce cas K est la transformée de Radon [42] qu'il s'agit d'inverser pour obtenir les informations 3D sur les centres diffuseurs du corps humain (y). Un dernier exemple en géophysique: après une explosion dans un milieu, on récupère les mesures à la surface x et l'on aimerait récupérer toutes les densités des roches dans le milieu (y) traversé par les ondes sismiques. Ce sont donc des ***problèmes inverses*** qui couvrent énormément de champs disciplinaires.

Soit $\hat{y}$ l'estimateur de y, on peut se dire que

$$\hat{y} = \underset{y}{\operatorname{argmax}}\ p(y|x) \tag{44}$$

c'est ce qu'on appelle le ***maximum a posteriori*** aussi nommé MAP qui est le résultat d'une approche typiquement bayésienne (nb. *a posteriori* car il faut d'abord obtenir la probabilité possiblement paramétrée). En utilisant la formule de Bayes:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{45}$$

où il est d'usage de nommer [43] $p(x|y)$ le ***likelihood***, $p(y)$ le ***prior*** sur y et $p(x)$ l'***évidence*** des mesures x. Ainsi, on peut écrire

$$\hat{y} = \underset{y}{\operatorname{argmax}}\ [p(x|y)p(y)] = \underset{y}{\operatorname{argmax}}\ [\log p(x|y) + \log p(y)] \tag{46}$$

Donc, il nous faut pouvoir estimer les deux quantités $\log p(x|y)$ (*log-likelihood*) et $\log p(y)$.

Si, l'on suppose que le bruit est *blanc gaussien* de variance σ^2 ($\eta \sim \mathcal{N}(0, \sigma^2)$), alors il est simple de conclure que

$$x - Ky \sim \exp\left(-\frac{\|x - Ky|^2}{2\sigma^2}\right) \Rightarrow \log p(x|y) = -\frac{\|x - Ky\|^2}{2\sigma^2} + cte \tag{47}$$

Quid alors du terme $\log p(y)$? Il s'agit de construire un modèle des données y. C'est là où ***la modélisation des données*** décrites dans les paragraphes précédents aide à la résolution du problème (nb. reste l'optimisation pour obtenir le maximum). Ceci dit avec

42. Johann Karl August Radon (1887-1956)

43. NDJE. Souvent on a affaire plutôt à des mesures x et des paramètres θ en lieu et place de y, et dans ce cadre le *prior* est celui des informations que l'on a *a priori* sur les dits paramètres, et $p(\theta|x)$ est la connaissance *a posteriori* sur θ une fois tenue compte des informations contenues dans les mesures x.

cette méthodologie, on peut appréhender beaucoup de problèmes de physique par exemple où l'hypothèse de bruit gaussien des mesures peut être une bonne approche en première intention.

3.4.2 Classification/régression

On peut également s'attaquer à des problèmes de classification/régression, où y est la classe recherchée sachant x le signal. Le classificateur de Bayes est de nouveau

$$\hat{y} = \underset{y}{\text{argmax}}\ p(y|x) = \underset{y}{\text{argmax}}\ \log p(x|y) = \underset{y}{\text{argmax}}\ [\log p(x|y) + \log p(y)] \tag{48}$$

Par contre, ce qui est "facile" dans ce cas de figure c'est d'obtenir $\log p(y)$, pourquoi? En effet, on peut avoir une approche fréquentiste pour obtenir *a priori* les probabilités de chaque classe, à partir de la base de données dont on dispose (ex. ImageNet). Par contre ce qui est difficile c'est d'estimer $p(x|y)$, comme par exemple quelle est $p(x)$ (x les pixels d'une image) sachant que l'on a que des images de la classe "*chien*"? C'est là encore où ***la modélisation de la probabilité*** $p(x|y)$ ***nous permet de résoudre ce problème***, mais encore faut-il avoir le bon modèle!

Ce qui est souvent utilisé, c'est une ***représentation*** $\Phi(x) = \{\phi(x)_k\}_{k\leq K}$ (*feature characteristics*) pertinente des données, puis on effectue une ***classification linéaire***[44] dite aussi *régression logistique* (Fig. 11):

$$\log p(x|y) = \sum_{k=1}^{K} \theta_k(y)\phi(x)_k + b = \Theta^T(y)\Phi(x) + b \tag{49}$$

qui est identique à l'expression Eq. 39. Et donc en sous-jacent, si l'on dispose des moments μ_k tels que

$$\mu_k = \mathbb{E}_{x\sim p(x|y)}[\phi_k(x)] \tag{50}$$

alors on aurait une caractérisation complète de $p(x|y)$ (les θ_k seraient les multiplicateurs de Lagrange).

Donc utiliser un classificateur linéaire, c'est une manière de dire que l'on a trouvé une représentation des données qui permet de caractériser $p(x|y)$. Typiquement, $\Phi(x)$ *est soit*

44. NDJE. Voir Sec. 7.3.3 du Cours de 2019, Sec. 5.2 du Cours 2020.

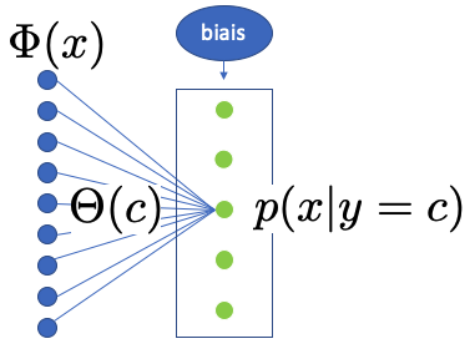


FIGURE 11 – Schématisation de la classification linéaire. La représentation $\Phi(x)$ est soit obtenue *a priori*, soit obtenue par apprentissage d'un réseau de neurone par exemple.

obtenue par des arguments a priori (ex. symétries/invariances, méthodes à noyaux, etc) *soit par l'apprentissage d'un réseau de neurones* qui comportent des filtres convolutionnels par exemple.

Ainsi, le commun des problèmes est la détermination de distributions de probabilités en grande dimension. Voyons ce qu'il en est en Physique Statistique. Cela va nous donner une base d'intuition sur ce que sont les concepts d'énergie, entropie, etc.

3.5 Physique Statistique: les deux premiers Principes

Nous allons visiter les deux Principes de la Thermodynamique qui mettent en jeu les deux notions d'énergie et d'entropie.

Concernant l'***énergie***, on peut dire que c'est une quantité, transmise à un système (ou bien interne à) sous forme de chaleur, de travail, d'interaction. On peut considérer, l'*énergie cinétique* liée au mouvement, les différents types d'*énergie potentielle* (c'est-à-dire qu'elles dérivent de potentiels, électrique, magnétique, gravitationnel...) qui est stockée en quelque sorte, et également l'*énergie de masse* en relativité. Le premier principe est un principe de conservation qui s'énonce ainsi

> **Premier Principe**: *Au cours d'une transformation quelconque d'un système fermé, la variation de son énergie est égale à la quantité d'énergie échangée avec le milieu extérieur, par transfert thermique (chaleur) et transfert mécanique (travail):*
>
> $$\Delta U = \Delta Q + W \tag{51}$$

et donc par conséquence si l'on considère l'***énergie totale***, du système et de l'extèrieur, alors celle-ci est ***conservée***; de même si le ***système est isolé du milieu extérieur (pas d'échange) alors son énergie est constante***. Si ***le système est isolé***, il peut y avoir des échanges entre les différents types d'énergie (ex. cinétique-potentielle) que l'on peut traduire par

$$\Delta E_c + \Delta E_p = \Delta U = 0 \tag{52}$$

Le lien avec ce qu'on a vu aux sections précédentes est dû à Maxwell et Boltzmann dans le cas de l'***étude de la cinétique des gaz***, gaz parfait de n particules de taille très petite par rapport à leur libre parcours moyen. Ce que montre Boltzmann (son théorème H) c'est que la distribution des vitesses au cours du temps va tendre vers une loi *stationnaire* de répartition qui est donnée par la distribution de Maxwell. Ce qui est curieux de prime abord (et à choquer si l'on peut dire) est que ***les lois des chocs considérés comme élastiques***[45] ***sont réversibles***, or on aboutit inexorablement, quelque soit la distribution initiale des vitesses, à ***une convergence vers la loi de Maxwell***. Autrement dit, ***on a un processus irréversible macroscopique produit par en sous-jacent des processus microscopiques réversibles***: comment est-ce possible? En fait, la création d'entropie est due à la sensibilité des chocs de deux particules qui étant donné leurs petites tailles peut induire brutalement des changements de directions (c'est une forme de chaos). C'est ce point qui a beaucoup fait couler d'encre.

3.5.1 Entropie et Irréversibilité

La notion d'irréversibilité est issue de l'idée que certains processus physiques empêchent des systèmes de revenir à leurs états initiaux: ex. typique la propagation de la

45. nb. comme les chocs de boules de billard, en opposition aux chocs inélastiques comme une boule qui tombe dans le sable.

chaleur et d'une manière générale les phénomènes de diffusion (échelle microscopique), mais aussi de convection (échelle macroscopique)... S. Carnot (1824), puis R. Clausius (1854) et L. Boltzmann (1872) vont fixer le cadre du second Principe de la Thermodynamique

> **Second Principe**: *toute transformation d'un système thermodynamique s'effectue avec augmentation de l'entropie globale incluant l'entropie du système et du milieu extérieur. On dit alors qu'il y a création d'entropie.*
>
> $$\Delta\mathbb{H}_{tot.} = \Delta\mathbb{H}_{syst.} + \Delta\mathbb{H}_{ext.} = \mathbb{H}_{crée} \geq 0 \tag{53}$$

(nb. il se peut que le système se structure par une transformation, et de ce fait $\mathbb{H}_{syst.}$ peut être négatif, mais alors cela se fait au détriment du désordre induit sur l'extérieur: ex. le processus de condensation de la vapeur en liquide provoque un dégagement de la chaleur). L'entropie du système, lors d'une transformation à température constante $T_{ext.}$ fixée par l'extérieur peut également se mettre sous la forme

$$\Delta\mathbb{H}_{syst.} = \mathbb{H}_{crée} - \Delta\mathbb{H}_{ext.} \geq -\Delta\mathbb{H}_{ext.} \Rightarrow \Delta\mathbb{H}_{syst.} \geq \frac{Q_{ext.\to syst.}}{T_{ext.}} \tag{54}$$

où $Q_{ext.\to syst.}$ est l'échange de chaleur cédée au système par l'extérieur. ***Les échanges réversibles se caractérisent par une égalité, alors que les échanges irréversibles produisent une inégalité stricte*** (résultat de Carnot). De plus, ***l'état d'équilibre*** consistant à ce qu'il n'y ait plus de création d'entropie $\mathbb{H}_{crée} = 0$, l'***entropie est alors maximum***. Nous n'entrons pas dans toutes les conséquences de ce principe tant du point de vue applicatif (ex. moteurs thermiques, cycle de Carnot, réactions chimiques, etc) que du point de vue philosophique (ex. autour de la notion de flèche du temps).

Nous verrons que la ***distribution de probabilité est stationnaire***, c'est-à-dire ne dépend plus du temps, et que ***tous les micro-états accessibles au système sont équiprobables***. Dans ces conditions, si l'on note Ω l'ensemble des configurations possibles accessibles au système alors

$$P(c \in \Omega) = \frac{1}{|\Omega|}, \qquad \mathbb{H} = k \log |\Omega| \qquad \text{(Boltzmann)} \tag{55}$$

Nous reviendrons sur ces notions lors des prochaines séances.

4. Séance du 1er Fév.

Durant cette séance, nous allons étudier l'***entropie probabiliste*** à la fois en ***Physique Statistique*** et en ***Théorie de l'Information***[46]. Tout d'abord, sous l'œil de ***Ludwig Boltzmann*** l'entropie est liée au ***nombre d'états accessibles au système*** (à l'équilibre). Nous verrons alors comment la pression, la température apparaissent comme des grandeurs macroscopiques définies à partie des variations de l'entropie. Puis, le point de vue de ***Claude Shannon*** (1916-2001) nous donnera une vision de l'entropie beaucoup plus générique, au sens où elle devient ***une notion purement mathématique***, qu'elle est complètement reliée aux ***distributions de probabilités*** (indépendamment du contexte où celles-ci sont employées), et qu'elle spécifie des ***phénomènes de concentration*** sur des ***ensembles typiques***. En sous-jacent de ces ensembles typiques on va retrouver la structure des ***ensembles micro-canoniques*** de Boltzmann. De plus la notion de concentration va nous donner un cadre de l'intuition selon laquelle les données (ex. images de...) s'agrègent sur des variétés. Nous verrons dans une autre séance, que l'évolution temporelle des distributions, si elle est décrite par une chaîne de Markov, nous donne le second principe de la thermodynamique dans lequel l'entropie augmente inexorablement (nous verrons qu'il y a une condition). Ainsi, les propriétés physiques élevées au rang de Principes à la manière de postulat), en quelque sorte émergent de propriétés mathématiques.

4.1 Point de vue de Boltzmann: ensemble micro-canonique

4.1.1 Entropie maximum, équiprobabilité des micros-états

Nous allons nous placer dans le cadre d'un système à l'***équilibre*** dont on fixe comme ***constantes*** les ***variables d'état*** suivantes: *volume* V, le *nombre* de corpuscules noté dans ce cours d, ainsi que l'*énergie* (interne) U. L'on considère alors la notion de ***micro-état*** constitué par l'ensemble des coordonnées et moments entrant dans la description de l'hamiltonien du système en mécanique classique. En mécanique quantique[47], les états ont

46. Voir également le cours de 2022.

47. nb. rappelons que L. Boltzmann n'a aucune idée qu'une telle théorie sera développée à peu près vingt ans après sa disparition.

des énergies quantifiées mais la philosophie reste inchangée. L'ensemble des *micro-états ou configurations* est noté Ω et $|\Omega|$ en est le cardinal.

NDJE. $|\Omega|$ est le nombre de configurations qui donnent des systèmes certes différents mais dont l'énergie se trouve dans l'intervalle $[U, U+\delta U]$, il est aussi appelé poids thermodynamique. Si l'on veut fixer les idées, il est donné par les expressions

$$|\Omega(U,d,V)| = \frac{1}{\prod_i h^{3N_i} N_i!} \int_{U \leq \mathscr{H} \leq U+\delta U} d\Gamma \qquad (classique) \tag{56}$$

$$= \sum_{U \leq U_c \leq U+\delta U} 1 \qquad (quantique) \tag{57}$$

avec $d\Gamma$ l'élément de l'espace de phase classique des degrés de liberté $\prod_i d^3q_i d^3p_i$ ($\mathscr{H}$ hamiltonien du système), et U_c une valeur propre de l'hamiltonien quantique.

La seconde loi de la Thermodynamique va nous dire qu'à l'équilibre la probabilité de chaque micro-état est une constante de valeur $P = 1/|\Omega|$. C'est ce que l'on appelle ***le principe fondamental d'équiprobabilité***. L. Boltzmann définit alors l'entropie selon [48]

$$\mathbb{H}(U,d,V) = k \log |\Omega(U,d,V)| = -k \log P \qquad (Boltzmann) \tag{58}$$

Comment L. Boltzmann élabore-t'il cette définition? En effet le lien avec la définition qui donne l'expression de l'entropie selon

$$\mathbb{H}(U,d,V) = -k \sum_{c \in \Omega} p_c \log p_c \qquad (Gibbs) \tag{59}$$

c'est-à-dire une formulation plus en lien avec les probabilités a été fait par J. W. Gibbs en 1901. En fait si l'on définit pour une distribution de probabilité quelconque (ici on se fixe des états discrets pour simplifier les notations), nous avons

> **Théorème 1** ***uniformité de la distribution de probabilité micro-canonique, maximum d'entropie***

48. nb. la constante k, dite de Boltzmann notée parfois k_B, vaut $k_\mathrm{B} = 1{,}380\,649 \times 10^{-23}\ \mathrm{J \cdot K^{-1}}$, elle est exacte et définit le Kelvin.

$$\sum_{c\in\Omega} p_c = 1, \qquad \mathbb{H}[p] = -\sum_{c\in\Omega} p_c \log p_c \qquad \Longrightarrow \qquad \max_p \mathbb{H}[p] = \log|\Omega| \tag{60}$$

Démonstration 1.
Nous utilisons la technique des multiplicateurs de Lagrange[49] sur laquelle nous reviendrons. Ainsi, on peut minimiser $-\mathbb{H}[p]$ en étudiant le point selle de la fonction

$$\mathscr{L}(\{p_c\},\lambda) = \sum_{c\in\Omega} p_c \log p_c + \lambda(\sum_{c\in\Omega} p_c - 1) \tag{61}$$

Ainsi $\forall i$

$$\frac{\partial \mathscr{L}}{\partial p_i} = \log p_i + 1 + \lambda = 0 \tag{62}$$

donc $\log p_i$ est une constante, il en est de même de p_i, et en utilisant la contrainte sur la somme des p_i, on conclut que

$$\forall c \in \Omega, \quad p_c = \frac{1}{|\Omega|} \Longrightarrow \mathbb{H}[p] = -\log|\Omega| \tag{63}$$

■

Donc, l'idée d'une ***distribution uniforme*** sur l'espace des configurations correspond bien au principe d'***entropie maximum***.

4.1.2 Équilibre thermodynamique

Si la distribution de probabilité est indépendante du temps, le système est à l'équilibre, ce qui peut en soit être considéré comme une définition de la notion d'équilibre. Il y en a une autre, et donc nous avons les définitions équivalentes suivantes:

49. Voir aussi la Sec. 8.3 du Cours de 2018.

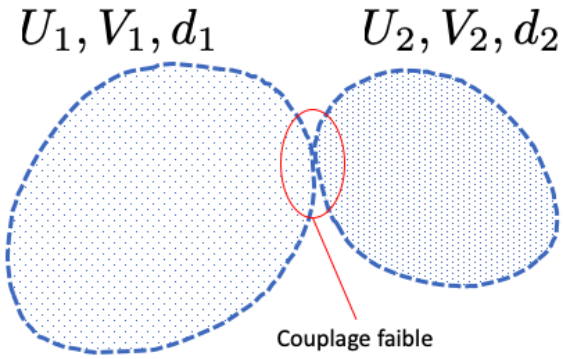


FIGURE 12 – Schéma de deux sous-systèmes en interaction faible.

Définition 1 ***(équilibre d'un système)***
Un système est à l'équilibre

a) si la distribution de probabilité des micros états est indépendante du temps;

b) ou si l'on considère des sous-systèmes de taille suffisamment grande pour leur associer les variables d'état, alors ils sont eux-mêmes à l'équilibre.

Examinons la définition (b) où l'on considère deux sous-systèmes d'un système isolé, identifiés par les triplets de valeurs d'état (U_1, V_1, d_1) et (U_2, V_2, d_2) (Fig. 12). On considère également que les deux systèmes sont faiblement couplés à savoir que les interactions à la surface d'échange sont à courte portée et négligeables devant les interactions entre corpuscules à l'intérieur des deux systèmes. De plus les fluctuations statistiques des variables d'état dues aux échanges sont négligeables. L'hypothèse d'isolement indique que les variables (V, d, U)

$$V = V_1 + V_2 \qquad d = d_1 + d_2 \qquad U = U_1 + U_2 \tag{64}$$

sont des constantes. Le nombre de configurations $|\Omega|$ du système constitué des deux sous-systèmes $i = \{1, 2\}$, sous l'hypothèse de faible couplage impliquant une indépendance, est donné par la relation

$$|\Omega_{1+2}(V, d, U)| = \prod_i |\Omega(V_i, d_i, U_i)| \tag{65}$$

Ainsi l'entropie du système globale $\mathbb{H}_{1+2}$ n'est autre que la somme des entropies des deux sous-systèmes, soit

$$\mathbb{H}_{1+2} = \mathbb{H}_1 + \mathbb{H}_2 \qquad (\textit{additivité de l'entropie}) \tag{66}$$

A l'équilibre du système globale, l'entropie est constante, donc $d\mathbb{H}_{1+2} = 0$.

Examinons le cas où $\{V_i, d_i\}_i = cte$, il vient

$$d\mathbb{H}_{1+2} = \sum_{i=1,2} \left(\frac{\partial \mathbb{H}_i}{\partial U_i}\right)_{V_i,d_i} dU_i \tag{67}$$

De plus $dU = 0$ indique que $dU_1 = -dU_2$, d'où

$$\left(\frac{\partial \mathbb{H}_1}{\partial U_1}\right)_{V_1,d_1} = \left(\frac{\partial \mathbb{H}_2}{\partial U_2}\right)_{V_2,d_2} \tag{68}$$

Or, ***la température d'un système à l'équilibre est reliée à l'entropie*** selon

$$\frac{1}{T} = \left(\frac{\partial \mathbb{H}}{\partial U}\right)_x \tag{69}$$

où x désigne l'ensemble des variables autres que U (ici V et d). Donc, à l'équilibre nous avons

$$T_1 = T_2 \qquad (\textit{équilibre}) \tag{70}$$

donc ***les températures de tous les sous-états sont identiques***.

Qui plus est à l'équilibre du système total, l'entropie $\mathbb{H}_{1+2}$ étant maximale, il vient

$$\left(\frac{\partial^2 \mathbb{H}}{\partial U^2}\right)_X \leq 0 \Longrightarrow \left(\frac{\partial T}{\partial U}\right)_X \geq 0 \tag{71}$$

donc à l'équilibre la température d'un sous-système croit avec son énergie interne. Ainsi, $T_1(U_1, X_1 = cte)$ et $T_2(U_2, X_2 = cte) = T_2(U - U_1, , X_2 = cte)$ évoluent en sens opposé lors de l'évolution de U_1: si U_1 croit par un transfert de $2 \to 1$, alors T_1 croit et T_2 décroit, ce qui tend à diminuer le transfert (si U_1 décroit, on peut raisonner sur U_2 qui croit). Donc, le ***transfert d'énergie va dans le sens d'homogénéiser les températures*** des deux sous

systèmes ce qui est également un résultat classique de thermodynamique.

4.1.3 Energie libre: volume variable

Relaxons le fait que les volumes $(V_i)_i$ soient constants, tandis que l'on garde constantes les variables x_i autres que les énergies et les volumes. Cependant, le système global est isolé, donc conserve ses variables (U, d, V) constantes. De ce fait, les différentielles totales des entropies respectives des deux sous systèmes s'écrivent alors

$$\forall i, \quad d\mathbb{H}_i = \left(\frac{\partial \mathbb{H}_i}{\partial U_i}\right)_{V_i, x_i} dU_i + \left(\frac{\partial \mathbb{H}_i}{\partial V_i}\right)_{U_i, x_i} dV_i \tag{72}$$

Mais $dV_1 = -dV_2$ et $dU_1 = -dU_2$, et à l'équilibre $d\mathbb{H}_1 + d\mathbb{H}_2 = 0$ donc on en déduire que

$$\left(\frac{\partial \mathbb{H}_1}{\partial U_1}\right)_{V_1, x_1} = \left(\frac{\partial \mathbb{H}_2}{\partial U_2}\right)_{V_2, x_2} \qquad \left(\frac{\partial \mathbb{H}_1}{\partial V_1}\right)_{U_1, x_1} = \left(\frac{\partial \mathbb{H}_2}{\partial V_2}\right)_{U_2, x_2} \tag{73}$$

La première relation redonne le résultat que ***la température des deux sous systèmes sont égales*** $(T_1 = T_2)$. En utilisant la définition de la pression P selon

$$P = T\left(\frac{\partial \mathbb{H}}{\partial V}\right)_{U, x} \tag{74}$$

avec x les autres variables, la seconde relation donne lieu à l'***égalité des pressions*** à l'équilibre

$$P_1 = P_2 \qquad (\textit{équilibre}) \tag{75}$$

Notons que la pression est une variable *extensive* comme la masse, le nombre de corpuscules, l'énergie, tandis que la température est une variable *intensive* comme la masse volumique, les concentrations. L'égalité des pressions induit qu'à l'interface entre les deux sous systèmes qu'il n'y a pas de travail d'une quelconque force.

D'une manière générale à l'équilibre à P et T fixée, la variation d'entropie est donnée par

$$d\mathbb{H} = \frac{1}{T} dU + \frac{P}{T} dV \tag{76}$$

Ainsi, ***la variation d'entropie a donc deux contributions*** (dans notre cas) l'une liée à la variation d'énergie interne, et l'autre liée à la variation d'une autre grandeur, ici le volume et le coefficient de variation défini une variable associée, ici la pression, normalisée par la température. Cette seconde contribution n'est autre que l'opposé du travail extérieur [50] sur le système W_{ext}. Or, ceci est à mettre en correspondance avec l'équation 41, où l'on identifierait dans ce cas l'***énergie libre*** comme la contribution du travail extérieur lié à la variation de volume. Ces deux composantes des variations de l'entropie montre un caractère général:

- une composante qui en thermodynamique est reliée à l'énergie interne que l'on peut se représenter au niveau microscopique comme l'agitation cinétique des corpuscules, et qu'au niveau des probabilité on représente comme $U_\theta(x) = \Theta^T\Phi(x)$;
- et une contribution qui apparait en thermodynamique ici comme un travail externe, et en probabilité apparait à travers la constante de (re)normalisation, c'est-à-dire $F_\theta = -\log Z_\theta$.

D'où une sorte d'analogie:

Thermo.	Stat.
Energie interne (cinétique)	$U_\theta = \Theta^T\Phi(x)$
Energie libre (travail ext.)	$F_\theta = -\log Z_\theta$
$1/T$, P/T, ...	multiplicateurs de Lagrange

Ce qu'il faut retenir, c'est que ***l'entropie est La Fonction centrale, et sa maximisation décrit le système à l'équilibre***. La question qui va nous occuper est de comprendre pourquoi cette notion émerge, d'où vient-elle? Il y a quelque chose de fondamental en sous-jacent: les fluctuations macroscopiques des variables d'état vont pouvoir être négligées et vont converger vers des constantes, or ceci se fait grâce à la ***loi des grands nombres qui donne le phénomène de concentration***, mais à la base il y a la notion d'***indépendance*** (ou au moins faible corrélation).

50. nb. la force de pression est orientée vers l'extérieur du volume si P est sa pression interne.

4.2 Information et codage

4.2.1 La loi des grands nombres

Soit A_d une *v.a* qui dépend de d, le nombre d'observations, telle que $A_d \xrightarrow[d\to\infty]{} A$, alors la convergence en probabilité nous dit que

$$\left(\forall \varepsilon > 0,\ \lim_{d\to\infty} \mathbb{P}\left[|A_d - A| \leq \varepsilon\right] = 1\right) \Leftrightarrow \left(A_d \xrightarrow[d\to\infty]{prob.} A\right) \tag{77}$$

En quelque sorte cela nous dit qu'il est *rare* que A_d se démarque de sa valeur limite A.

Voici alors le théorème dans le cas où la *v.a*, notée $\bar{X}_d$ est la moyenne d'ensemble de d *v.a iid*:

Théorème 2 ***(loi faible des grands nombres)***
Soit des variables aléatoires $(X_i)_{i\leq d}$ iid, et $\mathbb{E}[X_i] = \mu < \infty$ alors si $\bar{X}_d = \frac{1}{d}\sum_{i=1}^{d} X_i$, on a une convergence en probabilité

$$\bar{X}_d \xrightarrow[d\to\infty]{prob.} \mu \tag{78}$$

La démonstration peut être retrouvée dans la section 3.3 du cours de 2022 basée sur l'inégalité de Bienaymé-Tchebychev dans le cas où $\mathbb{E}[X_i^2] < \infty$.

Donc on a un phénomène de concentration dès lors que l'on a affaire à des variables *iid*. Mais pourquoi l'entropie entre-t-elle en jeu? Là il faut se référer à l'article de 1948 de Cl. Shannon *"A Mathematical Theory of Communication"*[51].

4.2.2 Indépendance et concentration (cas discret)

Quand on considère la Théorie de l'Information il y a deux regards nous dit S. Mallat: un regard probabiliste qui est sans doute le plus efficace, mais on peut également avoir une

51. C. E. Shannon, The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October. https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf.

vision déterministe sur la notion de codage en adoptant le point de vue de Kolmogorov [52]. Kolmogorov définit la quantité d'information à travers la taille de l'algorithme, exécuté sur machine de Turing, qui est capable de reproduire une séquence [53]. Les deux notions sont équivalentes, essentiellement du fait de phénomènes de concentration. Cependant, il est très difficile de faire des calculs en suivant le schéma de Kolmogorov, alors que le point de vue de Shannon est fait pour cela.

Soit $\{y_k\}_{k\leq K}$ un alphabet à K symboles que l'on note $\mathcal{A}$. Soit $(X_i)_i$ des *v.a iid* prenant leurs valeurs dans $\mathcal{A}$, on pose alors

$$\forall i,\ \mathbb{P}(X_i = y_k) = p(y_k) \tag{79}$$

tandis que l'entropie est définie selon

$$\mathbb{H} = -\sum_{y_k\in\Omega} p(y_k)\log p(y_k) \tag{80}$$

Comme les probabilités $p(y_k) \in [0,1]$ et que l'entropie est maximale quand toutes les probabilités $p(y_k)$ sont égales à $1/|\Omega| = 1/K$, alors on peut écrire

$$0 \leq \mathbb{H} \leq \log K \tag{81}$$

Or, si l'on considère le cas où $p(y_k) = 1$ pour $k = k_0$ et 0 si $k \neq k_0$, l'entropie est alors nulle ($\mathbb{H} = 0$); par contre si $\forall k$, $p(y_k) = 1/K$ alors $\mathbb{H} = \log K$. On sent bien que l'entropie est reliée à la ***notion d'incertitude***: il n'y a pas de doute dans le premier cas car toutes les *v.a* X_i prennent la valeur y_{k_0}, par contre l'incertitude est maximale dans le second cas (Fig. 13).

Examinons la probabilité jointe des événements lors de tirages des X_i. Comme ces variables sont *iid*, pour 1 tirage:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d) = \prod_{i=1}^{d} \mathbb{P}(X_i = x_i) \tag{82}$$

52. NDJE: Andrey N. Kolmogorov dés 1933 suivant les travaux d'Emile Borel (1871-1956) et d'Henri Lebesgue (1875-1941) élabore *la théorie des probabilités*, et établit un lien entre *mesure* et la *probabilité* des événements composés.

53. ex. π a une information au sens de Kolmogorov étant définit à travers une série entière qui peut être codée et exécutée sur un calculateur.

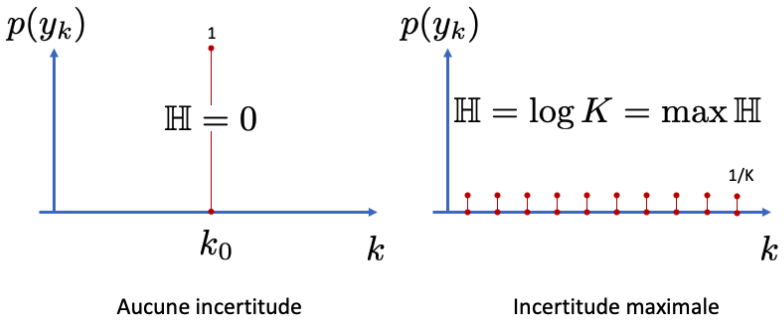


FIGURE 13 – Deux cas de figures extrêmes de distributions de probabilités qui illustrent la mesure de l'incertitude par la valeur de l'entropie.

Notons qu'ici on obtient 1 nombre, par contre si l'on procède à plusieurs tirages, alors $\mathbb{P}(X_1, X_2, \ldots, X_d)$ est elle-même une variable aléatoire. Cependant, l'indépendance nous donne la même décomposition

$$\mathbb{P}(X_1, X_2, \ldots, X_d) = \prod_{i=1}^{d} \mathbb{P}(X_i) \tag{83}$$

Ainsi en appliquant la loi des grands nombres

$$-\frac{1}{d} \log \mathbb{P}(X_1, X_2, \ldots, X_d) = \frac{1}{d} \sum_{i=1}^{d} (-\log \mathbb{P}(X_i)) \xrightarrow[d\to\infty]{prob.} \mathbb{E}[-\log \mathbb{P}(X_i)] = \mathbb{H}[X] \tag{84}$$

(nb. on utilise $\mathbb{H}[X]$ pour signifier que c'est l'entropie d'1 variable quelconque X_i). Donc, ***l'entropie apparait naturellement car on a affaire à des événements indépendants***. Considérons le corolaire suivant

Théorème 3 *Si les (X_i) sont iid de loi $\mathbb{P}(X)$, et prenant leurs valeurs dans $\mathcal{A} = \{y_k\}_{k \leq K}$ associé aux probabilités $p(y_k)$, alors*

$$-\frac{1}{d} \log \mathbb{P}(X_1, X_2, \ldots, X_d) \xrightarrow[d\to\infty]{prob.} \mathbb{H}[X] = -\sum_{k=1}^{K} p(y_k) \log p(y_k) \tag{85}$$

On peut réécrire la convergence en probabilité selon

$$\forall \varepsilon > 0, \lim_{d\to\infty} \mathbb{P}\left[\left|-\frac{1}{d}\log \mathbb{P}(X_1, X_2, \ldots, X_d) - \mathbb{H}[X]\right| \leq \varepsilon\right] = 1 \tag{86}$$

c'est-à-dire que pour d suffisamment grand

$$\mathbb{P}\left[\left|\underbrace{-\frac{1}{d}\log \mathbb{P}(X_1, X_2, \ldots, X_d)}_{f(\{X_i\}_{i\leq d})} - \mathbb{H}[X]\right| \leq \varepsilon\right] \geq 1 - \varepsilon \tag{87}$$

Cela nous dit que ***la fonction à*** d ***variables*** $f(\{X_i\}_{i\leq d})$ ***va se concentrer sur une surface définie par l'équation***

$$f(\{X_i\}_{i\leq d}) \approx \mathbb{H}[X] \tag{88}$$

avec une petite épaisseur ε. On voit poindre le paralèlle avec l'intuition que les images de chiens/chats s'agrègent sur une surface de $\mathbb{R}^d$. Pour aller plus avant, Cl. Shannon développe la notion d'ensemble *typique*.

4.2.3 Ensemble typique

D'après ce qui précède, on peut s'intéresser aux observations qui vont effectivement (pour un ε fixé) avoir une espérance qui va se retrouver à une distance ε de l'entropie. En notant, $\{X_i\}_{1\leq i\leq d} = \{x\}$, soit l'ensemble [54]

$$T_d^\varepsilon = \left\{\{x\} \in \mathcal{A}^d, \left|-\frac{1}{d}\log \mathbb{P}(\{x\}) - \mathbb{H}[X]\right| \leq \varepsilon\right\} \tag{89}$$

Notons que la condition d'appartenance à T_d^ε peut se mettre sous la forme (en prenant des logarithmes en base 2):

$$2^{-d(\mathbb{H}[X]+\varepsilon)} \leq \mathbb{P}(\{x\}) \leq 2^{-d(\mathbb{H}[X]-\varepsilon)} \tag{90}$$

c'est-à-dire qu'à un ε prés qui n'est pas totalement négligeable (cela gouverne la conver-

54. NDJE. d'un point de vue notation, il peut se faire qu'il y ait une confusion entre les variables d utilisée précédemment dans le cours, et n qui est généralement utilisée dans ce contexte et par exemple dans le cours de 2022.

gence de la concentration), la probabilité $\mathbb{P}(\{x\})$ est une constante donnée par

$$\mathbb{P}(\{x\}) \approx 2^{-d\mathbb{H}[X]} \tag{91}$$

Cela dit qu'***à l'intérieur de l'ensemble typique les probabilités sont quasiment uniformes.*** Notons que par additivité de l'entropie $\mathbb{H}[\{x\}] = d\mathbb{H}[X]$ ce qui permet de comprendre le scaling.

De plus nous avons les propriétés suivantes. Primo, pour tout $\varepsilon > 0$, et d assez grand selon Eq. 87

$$\mathbb{P}[\{x\} \in T_d^\varepsilon] \geq 1 - \varepsilon \tag{92}$$

c'est-à-dire que ***presque toutes les réalisations vont appartenir à T_d^ε, d'où l'appellation d'ensemble typique***. Secundo, le cardinal de l'ensemble typique satisfait [55]

$$(1-\varepsilon)2^{d(\mathbb{H}[X]-\varepsilon)} \leq |T_d^\varepsilon| \leq 2^{d(\mathbb{H}[X]+\varepsilon)} \tag{93}$$

ce qui donne par conséquent une relation

$$\mathbb{P}(\{x\}) \approx \frac{1}{|T_d^\varepsilon|} \tag{94}$$

Cette relation fait écho à la relation similaire en Physique Statistique (Th. 1), ce qui confère ***une relation étroite entre ensemble micro-canonique Ω et ensemble typique T_d^ε***. Les relations Eqs. 92,93 forment la base du fameux ***principe fondamental d'équi-probabilité*** que l'on définit comme une conséquence d'un phénomène de concentration des probabilités pour les ensembles typiques (***théorème d'équipartition asymptotique***). L'entropie donne le nombre d'états/configurations possibles qui ont une probabilité non négligeable même s'il elle peut être très faible (nb. $|\Omega|$ peut être très très grand).

4.2.4 Codage (typique)

Une autre notion que nous revisitons cette année concerne le *codage*, et en quoi l'entropie nous sert-elle dans ce contexte? Soit $X \in \mathcal{A}^d$, il peut se trouver soit dans

55. NDJE. La démonstration se trouve dans le cours de 2022 Sec. 6.4.

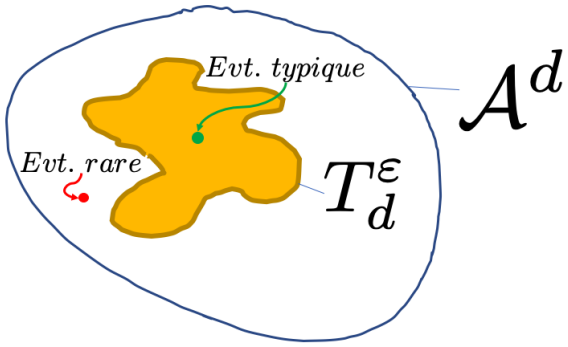


FIGURE 14 – Illustration d'un ensemble typique.

l'ensemble typique T_d^ε avec une "forte" probabilité, soit à l'extérieur avec une bien plus faible probabilité (Fig. 14). Donc, on peut penser au code suivant (dit ***code typique***):

- soit 1 bit qui code la qualité $X \in T_d^\varepsilon$ ou $X \notin T_d^\varepsilon$;
- si $X \in T_d^\varepsilon$, alors il peut être n'importe où, donc le code doit identifier tous les membres de l'ensemble. Sa longueur est donc $\ell(X) = \lceil \log_2 |T_d^\varepsilon| \rceil = \lceil d(\mathbb{H}[X] + \varepsilon) \rceil$;
- soit $X \notin T_d^\varepsilon$, on se dit que chacune des d composantes prend ses valeurs dans $\mathcal{A}$ de taille K, donc il nous faut un code de longueur $\ell(X) = \left\lceil \log_2 |\mathcal{A}^d| \right\rceil = \lceil d \log_2 K \rceil$.

En quelque sorte, ***l'idée est de prendre des codes plus courts quand la probabilité est grande et plus long quand la probabilité est petite***. Ce que l'on va optimiser, c'est le nombre de bits moyen. D'où le théorème suivant:

Théorème 4 *(**Borne de Shannon**)*
Le nombre de bits moyen par composante de $X \in \mathcal{A}^d$ d'un codage typique est tel que

$$R = \frac{1}{d} \sum_{X \in \mathcal{A}^d} \ell(X) \mathbb{P}(X) \leq \mathbb{H}[X] + C\varepsilon \tag{95}$$

Démonstration 4. Si pour un X quelconque, l'on décompose l'appartenance à l'ensemble

typique T_d^ε ou non, on a

$$\begin{aligned} R &= \frac{1}{d}\sum_{X\in T_d^\varepsilon}\ell(X)\mathbb{P}(X) + \frac{1}{d}\sum_{X\notin T_d^\varepsilon}\ell(X)\mathbb{P}(X) \\ &= \frac{1}{d}(\lceil d(\mathbb{H}[X]+\varepsilon)\rceil+1)\left(\sum_{X\in T_d^\varepsilon}\mathbb{P}(X)\right) + \frac{1}{d}(\lceil d\log_2 K\rceil+1)\left(\sum_{X\notin T_d^\varepsilon}\mathbb{P}(X)\right) \end{aligned} \tag{96}$$

Or $\sum_{X\in T_d^\varepsilon}\mathbb{P}(X)\leq 1$ et $\sum_{X\notin T_d^\varepsilon}\mathbb{P}(X)\leq\varepsilon$, de plus $\lceil x\rceil = \lfloor x\rfloor + 1 \leq x+1$ d'où

$$R \leq \frac{1}{d}(d(\mathbb{H}[X]+\varepsilon)+2) + \frac{1}{d}(d\log_2 K+2)\varepsilon \leq \mathbb{H}[X] + \varepsilon C' + \frac{2}{d} \tag{97}$$

ce qui pour d suffisamment grand fournit le résultat [56]. ■

Donc, ***le nombre de bits moyen par symbole du code typique est de l'ordre de l'entropie de la distribution de probabilité de chaque symbole***. En fait, on a une borne supérieure, et la question qui vient à l'esprit est de savoir si l'on peut faire mieux? L'espoir est de pouvoir trouver des ensembles plus efficaces que les ensembles typiques, car sinon le codage de Shannon est ce que l'on peut faire de mieux.

Donc, peut-on trouver de tels ensembles plus efficaces, ou dit autrement, ***des ensembles qui concentrent encore plus la probabilité sur un plus petit nombre d'événements*** que ce qui est donné par l'entropie.

Théorème 5 *(**optimalité des ens. typiques**)*
Soit $X = (x_1,\ldots,x_d)$ où les x_i sont des v.a iid de loi $p(x)$ (ex. les valeurs de x_i appartiennent à $\mathcal{A}$ de taille K). Soit B_δ^d le plus petit ensemble tq.

$$\mathbb{P}(X\in B_\delta^d) \geq 1-\delta \tag{98}$$

56. NDJE. une coquille c'était installée dans la démonstration de 2022 sans conséquence, mais donc je rectifie cette année.

alors $\forall \delta, \delta' > 0$,

$$\mathbb{P}(X \in B_\delta^d) \geq 1 - \delta \Rightarrow \frac{1}{d} \log_2 |B_\delta^d| \geq \mathbb{H}[X] - \delta' \tag{99}$$

La démonstration donnée dans le cours de 2022 Sec. 6.6 réside dans l'analyse de l'intersection entre B_δ^d et T_d^ε. On montre que les deux ensembles sont de tailles identiques (la probabilité de l'intersection est quasi égale à 1). ***Donc, l'ensemble de taille minimum qui concentre le lieu des observations est bien l'ensemble dont la taille est donnée par l'entropie, c'est-à-dire l'ensemble typique***. On ne peut donc trouver un codage qui dépasse le codage typique, et ***la borne donnée par le théorème 4 est optimale***. A la base du raisonnement, on a ***la loi des grands nombres*** qui nous donne la relation 84, laquelle est elle-même le fruit de ***l'indépendance*** des X_i.

Maintenant, le code typique n'est pas pratique car il faut savoir si l'élément appartient ou non à l'ensemble typique. Nous avions vu dans le cours de 2022, ***les codages entropiques instantanés*** (Shannon, Huffmann) [57]. Ce qui va nous occuper plutôt cette année c'est d'***étendre ces résultats au cas continu***, c'est-à-dire les x_i à valeurs dans $\mathbb{R}$ et non dans un alphabet discret (nb. les intensités d'un pixel d'une image, ou bien les échantillons d'un son sont en général numérisés mais nous pourrons imaginer que le pas de quantification est très petit devant la gamme dynamique).

5. Séance du 8 Fév.

Rappelons ce que l'on a vu, l'***entropie est une notion mathématique*** qui s'est affranchie de son cadre originel (cf. la Physique Statistique) grâce aux travaux de Cl. Shannon qui en fait une propriété intrinsèque des ***probabilités***. Surtout en grande dimension, elle aide à comprendre les ***phénomènes de concentration*** qui sont profondément reliés à la notion d'***indépendance statistique***. Or, nous devrons aborder le cas plus en rapport avec la réalité des données, à savoir qu'***il y a de la structure*** dès que l'on observe des images d'objets/scènes, des trames de musique/parole, texte, etc. Donc, nous allons voir comment la notion d'entropie des sections précédentes peut se ***généraliser dans le cas de variables***

57. Voir par ailleurs quelques mises en oeuvre https://github.com/jecampagne/cours_mallat_cdf/tree/main/cours2022

non-indépendantes, et de voir s'il y a malgré tout des phénomènes de concentration. On verra que c'est en effet le cas moyennant des hypothèses supplémentaires de ***stationnarité et d'ergodicité***. L'évolution temporelle de l'entropie sera vue dans le cadre markovien.

5.1 Entropie différentielle

Dans la séance précédente nous avons vu l'entropie dans le cas de variables prenant des valeurs dans un alphabet discret. Elle se généralise dans le cas de valeurs continues. Soit la définition suivante [58]

Définition 2 *(**Entropie différentielle**)*
Soit X v.a dont la probabilité de densité par rapport à la mesure de Lebesgue dx est notée $p(x)$ ($x \in \mathbb{R}$ ou $\mathbb{R}^n$). L'entropie différentielle est alors

$$\mathbb{H}_d = \mathbb{E}_{x\sim p(x)}[-\log p(x)] = -\int_{\mathbb{R}^n} p(x)\log p(x)\ dx \tag{100}$$

Contrairement à son équivalent "discret", l'entropie différentielle n'est pas forcément positive. L'exemple classique est de considérer une loi uniforme 1D, $x \sim \mathcal{U}[0,a]$, l'entropie différentielle est alors $\mathbb{H}_d = \log a$ qui devient négative dès lors que $a < 1$.

Un autre exemple plus intéressant est celui des lois gaussiennes:

$$X \sim \mathcal{N}(\mu,\sigma^2) \Rightarrow \mathbb{H}_d = \frac{1}{2} + \frac{1}{2}\log\Big(2\pi\sigma^2\Big) = \frac{1}{2}\log(2\pi e) + \log\sigma \tag{101}$$

Donc, quand $\sigma \ll 1$ alors $\mathbb{H}_d \ll 0$. Notons au passage que le facteur "$\log\sigma$" reflète un caractère général du rôle d'un facteur d'échelle, en effet

$$\forall\alpha > 0, \quad \mathbb{H}_d(\alpha X) = \mathbb{H}_d(X) + \log\alpha \tag{102}$$

qui vient du fait que $\frac{1}{\alpha}p(\frac{x}{\alpha}) = p(x)$. Quand on augmente la variance, on augmente l'incertitude sur la variable, donc on augmente l'entropie. Notez que l'intervalle de mesure donne le caractère relatif à la définition de l'entropie.

58. NDJE. J'opte pour le changement de notation opéré par S. Mallat de d à n pour la dimension de l'espace, sachant que d ici est relatif à *différentiel*. De plus ainsi, on garde les notations du Cours de 2022.

Maintenant avec ce nouvel outil, nous allons revisiter les propriétés vues dans le cas discret.

5.2 Équipartition asymptotique cas continu, ensemble typique

Si l'on se place de nouveau dans le cas de variables aléatoires indépendantes, alors on a le théorème suivant à rapprocher du Th. 3 vu dans le cas "discret":

Théorème 6 *Si les* $= (X_i)_{i\leq n}$ *sont* n *v.a iid prenant leurs valeurs dans* $\mathbb{R}$ $(\forall i, X_i \sim p)$ *alors*

$$-\frac{1}{n}\log p(x_1,\ldots,x_n) = -\frac{1}{n}\sum_{i=1}^{n}\log p(x_i) \xrightarrow[n\to\infty]{prob.} \mathbb{H}_d(x) = \mathbb{E}_{x\sim p(x)}[-\log p(x)]$$
$$= -\int_{\mathbb{R}} p(x)\log p(x)dx \quad (103)$$

et donc [59]

$$\forall \varepsilon \geq 0, \quad \mathbb{P}\left(\left|-\frac{1}{n}\log p(x_1,\ldots,x_n) - \mathbb{H}_d(x)\right| \leq \varepsilon\right) \geq 1-\varepsilon \quad (104)$$

Cette propriété nous conduit naturellement à la définition des ensembles typiques en tout point identique à celle du cadre discret (Sec. 4.2.3) $\{X\} = (x_1,\ldots,x_n)$

$$T_n^{\varepsilon} = \left\{\{X\} \in \mathbb{R}^n, \left|-\frac{1}{d}\log p(\{X\}) - \mathbb{H}_d(x)\right| \leq \varepsilon\right\} \quad (105)$$

alors on est quasiment sûr que $\{X\}$ est élément de l'ensemble typique

$$\mathbb{P}(\{X\} \in T_n^{\varepsilon}) \geq 1-\varepsilon \quad (106)$$

et nous allons retrouver les deux propriétés vues dans le cas discret.

Tout d'abord, l'expression 105 nous indique que

$$2^{-n(\mathbb{H}_d(x)+\varepsilon)} \leq p(\{X\}) \leq 2^{-n(\mathbb{H}_d(x)-\varepsilon)} \quad (107)$$

59. NDJE. petite erreur au tableau que tout le monde aura rectifiée.

ce qui fournit qu'à un ε près ***l'uniformité de la probabilité sur l'ensemble typique***.

De plus, la taille de l'ensemble typique qui ne peut être définie en comptant les éléments, a néanmoins la propriété suivante:

Théorème 7 *(**Volume typique**)*
Soit le volume d'un ensemble Ω est relativement à la mesure de Lebesgue:

$$V(\Omega) := \int_{\Omega} dx$$

Pour n assez grand

$$(1-\varepsilon)2^{n(\mathbb{H}_d(x)-\varepsilon)} \leq V(T_n^{\varepsilon}) \leq 2^{n(\mathbb{H}_d(x)+\varepsilon)} \tag{108}$$

Démonstration 7.

Dans un premier temps

$$1 = \mathbb{P}(\{X\} \in \mathbb{R}^n) \geq \mathbb{P}(\{X\} \in T_n^{\varepsilon}) = \int_{T_n^{\varepsilon}} p(x)dx \geq 2^{-n(\mathbb{H}_d(x)+\varepsilon)} \int_{T_n^{\varepsilon}} dx \tag{109}$$

d'où

$$V(T_n^{\varepsilon}) \leq 2^{n(\mathbb{H}_d(x)+\varepsilon)} \tag{110}$$

Dans un second temps

$$1-\varepsilon \leq \mathbb{P}(\{X\} \in T_n^{\varepsilon}) = \int_{T_n^{\varepsilon}} p(x)dx \leq 2^{-n(\mathbb{H}_d)(x)-\varepsilon} \int_{T_n^{\varepsilon}} dx \tag{111}$$

d'où

$$V(T_n^{\varepsilon}) \geq (1-\varepsilon)2^{n(\mathbb{H}_d(x)-\varepsilon)} \tag{112}$$

■

On retrouve également qu'à peu de chose près que la probabilité au sein d'un ensemble typique est l'inverse de sa taille:

$$\mathbb{P}(\{X\} \in T_n^{\varepsilon}) \approx \frac{1}{V(T_n^{\varepsilon})} \tag{113}$$

Ce qu'il faut en retenir est que tout repose sur la propriété de ***convergence en probabilité*** basée elle-même sur l'***indépendance*** des $v.a$ (X_i). On est certain que les données en grande dimension vont se retrouver dans T_n^ε avec une probabilité uniforme gouvernée par l'entropie différentielle. Ainsi se dessine le support de l'*intuition déterministe, dans une approche probabiliste*. L'autre résultat que l'on ne va pas démontrer, est que l'***on ne peut pas faire mieux dans ce cadre d'indépendance des variables.***

Ce cadre comme on l'a déjà dit, il faut pouvoir en sortir car le plus clair du temps les données sont en interaction. Pour mémoire, pensez aux pixels d'une image d'un objet (vase, table,...): le contour de cet objet nécessairement connecte un certain nombre de pixels possiblement à grande échelle, et pour reconnaitre cet objet parmi une scène, il faut bien rendre compte de ces connexions/interactions. C'est là où les choses deviennent plus subtiles. Ayons comme guide que l'entropie est une mesure de l'incertitude sur la séquence de n variables aléatoires $(x_1, \ldots, x_n)$. Or, s'il y a des dépendances entre ces variables, toutes ne sont pas indépendantes et il y a de la redondance, donc il y a moins d'incertitude, donc l'entropie doit diminuer, d'où la taille des ensembles typiques diminue également. Ce qui est une bonne chose, par exemple si l'on procède à du codage [60]. Donc abordons la notion de dépendance.

5.3 Dépendance et entropie (jointe, conditionnelle, relative)

Nous allons utiliser les probabilités conditionnelles afin de définir une entropie. Le cas discret [61] va nous donner un cadre pour nous familiariser tout en permettant de manipuler des entropies positives. Le cadre continu demande de trouver la bonne définition à l'aide de la distance de Kullback-Leibler.

60. En aparté, S. Mallat nous dit que certes on ne peut coder un nombre réel avec un nombre fini de bits, mais si on accepte une petite erreur de quantification, alors on peut procéder comme dans le cas discret. Se sont des problèmes de *Sphere packing*. Vous pouvez vous reporter au Cours de 2022, Sec. 8. NDJE. Dans cette thématique de *Sphere packing*, notons au passage qu'une Médaille Fields en 2022 a été attribuée à la mathématicienne d'origine ukrainienne Maryna Viazovska: "*[She] is awarded the Fields Medal 2022 for the proof that the E8 lattice provides the densest packing of identical spheres in 8 dimensions, and further contributions to related extremal problems and interpolation problems in Fourier analysis.*"

61. NDJE. Voir cours de 2022 (Sec. 6.2)

Définition 3 *L'entropie jointe de deux v.a X, Y à valeurs dans dans un alphabet $\mathcal{A}$*

$$\mathbb{H}(X,Y) := -\mathbb{E}_{(x,y)\sim p}[\log p(X,Y)] = -\sum_{k,k'} p(X=a_k, Y=a_{k'}) \log p(X=a_k, Y=a_{k'}) \tag{114}$$

et l'entropie conditionnelle

$$\begin{aligned}\mathbb{H}(Y|X) &:= \sum_k p(X=a_k)\mathbb{H}(Y|X=a_k) \\ &= -\sum_{k,k'} p(X=a_k, Y=a_{k'}) \log p(Y=a_{k'}|X=a_k) \\ &= -\mathbb{E}_{(x,y)\sim p}[\log(Y|X)] \end{aligned} \tag{115}$$

Une propriété relie ces deux entropies selon

Propriété 1

$$\mathbb{H}(X,Y) = \mathbb{H}(X) + \mathbb{H}(Y|X) \tag{116}$$

Si les variables sont indépendantes, on retrouve alors la loi d'additivité des entropies. La démonstration (Cours 2022 Sec. 6.2) repose sur la formule de Bayes des probabilités conditionnelles En effet [62],

$$\begin{aligned}\mathbb{H}(X,Y) - \mathbb{H}(X) &= -\sum_{x,y} p(x,y)\log p(x,y) + \sum_x p(x)\log p(x) \\ &= -\sum_{x,y} p(x,y)\log p(x,y) + \sum_x \left(\sum_y p(x,y)\right)\log p(x) \\ &= -\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)} = -\sum_{x,y} p(x,y)\log p(y|x) \\ &= \mathbb{H}(Y|X) \end{aligned} \tag{117}$$

Introduisons à présent l'***entropie relative*** également appelée la ***divergence de Kullback-Leibler*** [63] est très utilisée en probabilité pour calculer l'adéquation entre deux distributions, d'où son appellation abusive de distance Kullback-Leibler (elle n'est pas symétrique)

62. Pour simplifier la notation $p(x,y) = p(X=x, Y=y)$ idem pour $p(x)$, $p(y|x)$.
63. Voir aussi Cours 2019 Sec. 7.2.3

Définition 4 *(**Kullback-Leibler**)*
Si le support[a] *de q inclut le support de p alors*

$$D_{KL}(p\|q) := \int_{\mathbb{R}^n} p(x) \log \frac{p(x)}{q(x)} dx = \mathbb{E}_{X\sim p}\left[\log \frac{p(X)}{q(X)}\right] \tag{118}$$

L'intégrale pouvant se transformer en somme si besoin.

a. Par convention on pose $0 \log 0 = 0$ et $0 \log(0/0) = 0$.

On peut voir $D_{KL}(p\|q)$ sous l'angle d'une mesure de l'inefficacité de codage. En effet, la taille du codage optimal de $(x_1, \ldots, x_n)$ est de l'ordre de l'entropie de la probabilité associée aux x_i. Si par exemple celle-ci et p, donc la taille moyenne du code est de l'ordre de $\mathbb{E}_p[-\log p]$, mais si on se trompe[64] et que l'on pense que $x \sim q$, alors on obtient une taille moyenne de $\mathbb{E}_p[-\log q]$ (cf. les x sont toujours en sous-jacent tirés selon p inconnue). Donc la différence entre les deux tailles n'est autre que $D_{KL}(p\|q)$.

Les propriétés importantes de $D_{KL}(p\|q)$ sont les suivantes:

- $D_{KL}(p\|q) \neq D_{KL}(q\|p)$ qui vient de la définition même qui montre que ce n'est pas une distance;
- $D_{KL}(p\|q) \geq 0$ et $D_{KL}(p\|q) = 0 \Leftrightarrow p = q$.

Pour démontrer la seconde propriété on peut se servir du théorème suivant

Théorème 8 *(**Inégalité de Jensen**)*
*Soit f une **fonction convexe** en dimension 1 (dérivée seconde positive ou nulle), alors pour toute v.a X*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]) \tag{119}$$

*et si f est **strictement convexe** (dérivée seconde strictement positive), on a égalité ssi alors la seule valeur prise par X est $\mathbb{E}[X]$.*

64. C'est un cas usuel car, on modélise des données sans être certain que le modèle reflète la réalité. C'est certes un débat philosophique mais pas uniquement.

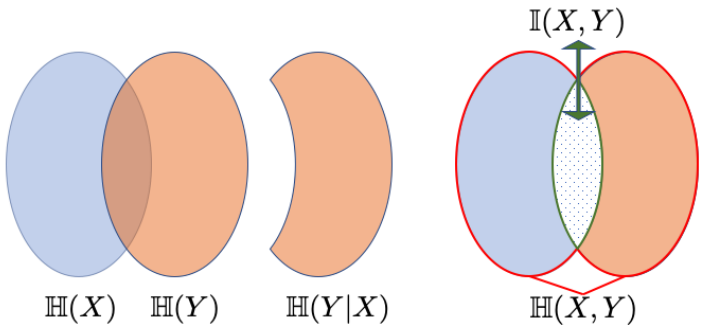


FIGURE 15 – Schématisation d'entropie $\mathbb{H}(X)$ et $\mathbb{H}(Y)$, de $\mathbb{H}(Y|X)$, ainsi que de l'information mutuelle et de l'entropie jointe.

Ainsi, comme la fonction log est strictement concave donc $-\log$ strictement convexe et

$$\begin{aligned} D_{KL}(p\|q) &= -\int p(x)\log\frac{q(x)}{p(x)}dx \\ &= \mathbb{E}_p\left[-\log\frac{q(x)}{p(x)}\right] \geq -\log\mathbb{E}_p\left[\frac{q(x)}{p(x)}\right] = -\log(1) = 0 \end{aligned} \tag{120}$$

La stricte concavité du log, nous dit que l'inégalité ci-dessus se transforme en égalité *ssi* $p(x)/q(x)$ prend une valeur unique. Soit c cette valeur, comme $\int p(x)dx = \int q(x)dx = 1$ alors $c = 1$, et donc on a le second résultat du théorème [65].

Afin d'appréhender la dépendance, utilisons la notion d'information mutuelle, mais avant cela nous pouvons démontrer la proposition suivante

Propriété 2

$$\mathbb{H}(X|Y) \leq \mathbb{H}(X) \tag{121}$$

ou autrement dit le conditionnement réduit l'incertitude.

65. NDJE. on aurait pu également se servir du fait que $\forall x > 0,\ -\log x \geq 1 - x$, l'égalité étant vraie *ssi* $x = 1$.

Ceci se démontre de la façon suivante

$$\begin{aligned}\Delta &= \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \\ &= \sum_{x,y} p(x,y)\log p(x,y) - \sum_{x} p(x)\log p(x) - \sum_{y} p(y)\log p(y) \\ &= \sum_{x,y} \left[p(x,y)\log p(x,y) - p(x,y)\log p(x) - p(x,y)\log p(y)\right] \\ &= \sum_{x,y} p(x,y)\log \frac{p(x,y)}{p(x)p(y)} = D_{KL}(p(x,y)\|p(x)p(y)) \geq 0 \end{aligned} \tag{122}$$

Au passage on s'aperçoit que l'égalité prévaut si les *v.a* X et Y sont indépendantes. Delà il devient naturel de définir l'information mutuelle (Shannon) selon

Définition 5 ***(Information mutuelle)***
Soit deux v.a X, Y *de probabilité jointe* $p(x,y)$ *et les marginales* $p(x)$, $p(y)$, *l'information mutuelle est la quantité suivante*

$$\mathbb{I}(X,Y) := D(p(x,y)\|p(x)p(y)) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \geq 0 \tag{123}$$

Si on a un conditionnement, il y a réduction d'entropie, ce qui se manifeste par $\mathbb{I}(X,Y) > 0$ qui quantifie en quelque sorte l'action qu'apporte la connaissance d'une variable sur une autre.

Un petit récapitulatif de ces notions sous forme de schéma est donné sur la figure 15. Maintenant, on peut s'attaquer au problème du calcul de l'entropie en grande dimension quand les variables ont des interdépendances.

5.4 Équipartition avec dépendance

5.4.1 Entropie moyenne, taux d'entropie

Quand on se place en dimension n, la définition de l'entropie ne pose pas de problème en soi. Soit n *v.a* $\{X\}_n = (X_1, \ldots, X_n)$ (ici à valeurs dans un alphabet)

$$\mathbb{H}(\{X\}_n) = -\mathbb{E}_{\{X\}_n \sim p}[\log p(\{X\}_n)] \tag{124}$$

Le point qui nous intéresse et comment cela va se comporter quand n tend vers l'infini. Or, nous avons l'intuition de la Physique Statistique où l'*entropie est une variable extensive* qui croit avec le nombre de particules. Donc, en étudiant la moyenne de $\mathbb{H}(\{X\})$ on pourrait s'attendre à une constante

Définition 6 ***taux d'entropie***

On appelle le taux d'entropie (entropy rate) la limite quand elle existe suivante:

$$\mathbb{H}(\chi) := \lim_{n\to\infty} \frac{1}{n}\mathbb{H}(\{X\}_n) \tag{125}$$

c'est l'entropie moyenne par élément/symbole.

Dans le cas *indépendant*, par l'additivité de l'entropie

$$\frac{1}{n}\mathbb{H}(\{X\}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{H}(X_i) \tag{126}$$

La limite existe-t-elle? Si elles ont même loi, alors la limite existe et nous donne l'équipartition asymptotique. Mais sinon, il n'y a aucune raison que la limite existe *a priori*. Prenons le cas où chaque X_i est une variable de Bernoulli de probabilité a_i pour 1 (a_i différent d'une *v.a* X_i à une autre).

$$\mathbb{H}[\mathscr{B}(a)] = -a\log(a) - (1-a)\log(1-a) = \begin{cases} 0 & a = 1 \\ 1 & a = 1/2 \end{cases} \tag{127}$$

On peut alors construire une série de X_i pour laquelle la moyenne oscille dans $[0,1]$ sans obtenir de limite [66].

Ceci dit à part le cas où toutes les variables X_i sont toutes indépendantes, la convergence va apparaître dans le cadre des ***processus stationnaires***. Cette hypothèse n'est pas restrictive bien au contraire: si avec une caméra CCD on prend une image, la loi de distribution des n *v.a* constituées par les intensités des n pixels est la même en gros quand

66. NDJE. Une tentative serait par exemple sachant que l'on peut faire en sorte que $\mathbb{H}[X_i] \in \{0,1\}$ alors les moyennes courantes sont éléments de $[0,1]$, alors on peut considérer la valeur de la moyenne à une étape k; puis compléter la suite des X_i $(i > k)$ avec des 1 pour faire approcher la moyenne aussi prés que l'on veuille de 1 (par valeur inf.), puis de compléter la dite série avec des 0 pour faire approcher la moyenne aussi prés que l'on veuille de 0 (par valeur sup.), et ainsi de suite.

on dirige la caméra ici ou là (cf. image et image translatée). Si on a une pixelisation de toute une scène et que l'on extrait des lots de n pixels translatés les uns par rapport aux autres alors l'hypothèse nous dit $p(X_1, \ldots, X_n) = p(X_{1+t}, \ldots, X_{n+t})$. ***Bien entendu, cette hypothèse est fausse dès lors que l'on procède par exemple à un recentrage des images*** de visages, chiffres, galaxies, etc. Donc attention, mais l'hypothèse de stationnarité liée à l'invariance par translation reste bien réalisée en pratique.

Entrons dans le détail et commençons par préciser la stationnarité

Définition 7 *(**stationnarité**)*
Un processus aléatoire est dit stationnaire ssi la probabilité jointe de toute séquence $X_1, X_2, \ldots, X_n$ est invariante par translation

$$\forall(n,k)\ \forall(x_1,\ldots,x_n), \quad p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = p(X_{1+k} = x_1, X_{2+k} = x_2, \ldots, X_{n+k} = x_n) \quad (128)$$

En particulier

$$\forall k, \quad p(X_1 = x_1) = p(X_{1+k} = x_1) \quad (129)$$

c'est-à-dire que la loi de X_1 est la même que n'importe quelle autre variable X_j, donc cela va assurer la convergence de la moyenne des entropies des X_i et donc l'existence de $\mathbb{H}(\chi)$. De même

$$p(X_1 = x_1 | X_2 = x_2) = \frac{p(X_1 = x_1, X_2 = x_2)}{p(X_2 = x_2)} = \frac{p(X_{1+k} = x_1, X_{2+k} = x_2)}{p(X_{2+k} = x_2)} = p(X_{1+k} = x_1 | X_{2+k} = x_2) \quad (130)$$

que l'on peut généraliser aux probabilités conditionnelles $p(X_n|X_1, \ldots, X_{n-1})$ par exemple.

Petit exemple, où les X_i ont des valeurs dans un alphabet de K symboles de probabilité uniforme $1/K$, l'hypothèse de stationnarité nous donne alors que $\mathbb{H}(\chi) = \log K$. Ceci dit dans une langue telle que le français la fréquence des lettres n'est pas uniforme, et la probabilité de deux lettres consécutives n'est pas égale non plus au produit des probabilités de chacune d'elle. Alors... c'est un autre cadre.

Cependant, nous allons définir une autre moyenne qui est plus maniable que cette entropie moyenne.

5.4.2 Entropie conditionnelle moyenne, taux d'entropie bis

Si on a une séquence de v.a $X_1, \ldots, X_n$ et que l'on veut la coder, une méthode est de le faire itérativement: on commence par coder X_1, puis on code X_2 sachant X_1, puis X_3 sachant X_2, X_1 et ainsi de suite. L'entropie conditionnelle $\mathbb{H}(X_n|X_1, \ldots, X_{n-1})$ fait appelle à probabilité conditionnelle. La question est de savoir si quand n tend vers l'infini, est-ce que cette entropie conditionnelle converge vers une limite? Il s'agira alors de l'information supplémentaire qu'apporte la variable X_n par rapport à toute l'information déjà récoltée avec les $n-1$ précédentes variables.

Théorème 9 *Dans le cas d'un processus stationnaire, primo* $\mathbb{H}(\chi)$ *existe et secundo*

$$\mathbb{H}'(\chi) := \lim_{n\to\infty} \mathbb{H}(X_n|X_1, \ldots, X_{n-1}) = \mathbb{H}(\chi) \tag{131}$$

Démonstration 9. La démonstration commence par faire remarquer que

$$\mathbb{H}(X_n|X_1, \ldots, X_{n-1}) \leq \mathbb{H}(X_{n-1}|X_1, \ldots, X_{n-2}) \tag{132}$$

En effet

$$\begin{aligned} \mathbb{H}(X_n|X_1, \ldots, X_{n-1}) &\leq \mathbb{H}(X_n|X_2, \ldots, X_{n-1}) && (Eq.\ 121) \\ &\leq \mathbb{H}(X_{n-1}|X_1, \ldots, X_{n-2}) && (Déf.\ 7) \end{aligned} \tag{133}$$

Donc, nous avons une suite de valeurs positives (cas discret) qui décroissent donc il y a convergence, et $\mathbb{H}'(\chi)$ existe. Ensuite,

$$\frac{1}{n}\mathbb{H}(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{H}(X_i|X_1 \ldots, X_{i-1}) \tag{134}$$

que l'on obtient par itération de Eq. 116. Si l'on prend la moyenne, le membre de gauche tend vers $\mathbb{H}(\chi)$. Pour le membre de droite, on peut utiliser le théorème de Cesàro:

Théorème 10 *(**somme de Cesàro**)*
Soit $(a_i)_{i>0}$ une suite de nombres ($\mathbb{R}$ ou même $\mathbb{C}$). Si la suite converge vers μ alors la suite des moyennes de terme général

$$b_n = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{135}$$

converge également et sa limite est μ.

Démonstration 10. La démonstration se fait comme suit. Primo la convergence de la suite des (a_i) nous permet de dire que $\forall \varepsilon \geq 0$, il existe N_ε tel que $\forall i \geq N_\varepsilon$, on ait $|a_i - \mu| \leq \varepsilon$.

Maintenant examinons $n > N_\varepsilon$:

$$\begin{aligned}
|b_n - \mu| &= \left|\frac{1}{n}\sum_{i=1}^{n}(a_i - \mu)\right| = \left|\frac{1}{n}\sum_{i=1}^{N_\varepsilon - 1}(a_i - \mu) + \frac{1}{n}\sum_{N_\varepsilon}^{n}(a_i - \mu)\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{N_\varepsilon}(a_i - \mu)\right| + \left|\frac{1}{n}\sum_{i=N_\varepsilon+1}^{n}(a_i - \mu)\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{N_\varepsilon}(a_i - \mu)\right| + \frac{1}{n}\sum_{i=N_\varepsilon+1}^{n}|a_i - \mu| \\
&\leq \frac{1}{n}\left|\sum_{i=1}^{N_\varepsilon}(a_i - \mu)\right| + \underbrace{\frac{n - N_\varepsilon}{n}\varepsilon}_{\leq \varepsilon}
\end{aligned} \tag{136}$$

La somme dans le premier terme de droite ne dépend pas de n, il existe donc un $N'_\varepsilon \geq N_\varepsilon$ tel que $\forall n \geq N'_\varepsilon$

$$\frac{1}{n}\left|\sum_{i=1}^{N_\varepsilon}(a_i - \mu)\right| \leq \varepsilon$$

donc également $\forall n \geq N'_\varepsilon$, $|b_n - \mu| \leq 2\varepsilon$, ce qui nous donne bien $\lim_{n\to\infty} b_n = \mu$ (nb. on peut faire en sorte que l'inégalité donne une borne ε au lieu de 2ε). ■

Ainsi on peut conclure que le membre de l'équation 134 converge vers $\mathbb{H}'(\chi)$, et ainsi on conclut à l'égalité

$$\mathbb{H}(\chi) = \mathbb{H}'(\chi)$$

■

Nous verrons comment ces résultats nous donnent accès à un ***théorème d'équipartition***, en particulier il va nous falloir rajouter une l'***hypothèse d'ergodicité***.

5.5 NDJE. Petit vadémécum dans le cadre continu

Je rassemble ici quelques définitions et relations dans le cas de *v.a* prenant leurs valeurs dans $\mathbb{R}$. Sur le thème d'entropie différentielle, nous avons introduit

$$\mathbb{H}_d(X) = -\int p(x)\log p(x)dx \qquad \textit{entropie diff.} \tag{137}$$

$$\mathbb{H}_d(X,Y) = -\iint p(x,y)\log p(x,y)dxdy \qquad \textit{entropie diff. conjointe} \tag{138}$$

$$\mathbb{H}_d(X|Y) = -\iint p(x,y)\log p(x|y)dxdy \qquad \textit{entropie diff. conditionnelle} \tag{139}$$

$$\mathbb{I}(X,Y) = D_{KL}(p(x,y)\|p(x)p(y)) \geq 0 \qquad \textit{information mutuelle} \tag{140}$$

avec les relations suivantes:

$$\mathbb{H}_d(X,Y) = \mathbb{H}_d(X|Y) + \mathbb{H}_d(Y) = \mathbb{H}_d(Y|X) + \mathbb{H}_d(X) \tag{141}$$

$$\mathbb{H}_d(X|Y) \leq \mathbb{H}_d(X) \tag{142}$$

$$\mathbb{I}(X,Y) = \mathbb{H}_d(X) + \mathbb{H}_d(Y) - \mathbb{H}_d(X,Y) \tag{143}$$

La plus part des relations ont été démontrées dans le cas de v.a à valeurs dans un alphabet discret, mais donc sont également valables dans le cas continu.

De même $X = (X_1, \ldots, X_n)$, on a les relations suivantes

$$\mathbb{H}_d(X+c) = \mathbb{H}_d(X) \qquad \forall c \in \mathbb{R} \tag{144}$$

$$\mathbb{H}_d(\alpha X) = \mathbb{H}_d(X) + \log|\alpha| \qquad \forall \alpha \neq 0 \tag{145}$$

$$\mathbb{H}_d(\mathbf{A}X) = \mathbb{H}_d(X) + \log|\det \mathbf{A}| \qquad \forall \mathbf{A} \in GL_n(\mathbb{R}) \tag{146}$$

6. Séance du 15 Fév.

NDJE. Je profite pour dire que j'ai mis en œuvre un repository GitHub pour quelques applications numériques illustrant le cours. https://colab.research.google.com/github/jecampagne/cours_mallat_cdf. Pour 2023, les notebooks pourront directement être exécutés sur Google Colab. La migration des nbs de 2022 est également prévue.

Nous allons explorer dans la continuité de la section précédente, l'exploration des systèmes plus complexes où les variables montrent des interdépendances, et nous rappelons que ce cas est bien celui qui nous concerne pour explorer les données structurées. L'ingrédient a avoir en tête est celui du ***taux d'entropie*** (Déf. 6) défini d'abord comme la limite pour $n \to \infty$ de la ***moyenne de l'entropie jointe*** de n *v.a* prenant leurs valeurs dans χ (ex. n pixels d'une image, ou n échantillons d'une trame sonore); et nous avons vu que pour des ***processus stationnaires***, on peut également la définir comme ***limite d'une suite d'entropies conditionnelles*** à travers le théorème 9.

Nous allons étudier les conséquences pour le codage et les phénomènes de concentration des données dans un ensemble typique. Ceci dit, l'hypothèse de stationnarité devra être complétée par celle d'***ergodicité*** ce qui nous fera visiter l'important ***théorème de Shannon–McMillan–Breiman*** qui donne la concentration, l'uniformité des probabilités sur les ensembles typiques dont les caractéristiques sont spécifiées par l'entropie. En pratique, nous étudierons l'exemple important des ***chaines de Markov*** typique d'un système physique où le temps est discrétisé et l'état est totalement spécifié par des variables à l'instant t sans mémoire. Nous verrons alors comment la notion d'***irréversibilité*** apparait dans ce contexte, et comment l'entropie augmente en fonction du temps ce qui constitue le ***2nd principe de la thermodynamique***.

6.1 Equipartition asymptotique avec dépendance: la condition

Dans le cas indépendant (Sec. 4.2.2) la probabilité jointe $\mathbb{P}(X_1, \ldots, X_n)$ est un produit, donc le logarithme une somme de n termes et la moyenne converge vers l'espérance du log de la probabilité d'où la concentration et l'apparition naturelle de l'entropie dans ce cas (Th. 3 (*cas discret*), 6 (*cas continu*)). Voyons à présent, dans le cas de dépendance comment ces notions apparaissent.

Il nous est toujours possible d'écrire

$$\mathbb{P}(X_1,\ldots,X_n) = \mathbb{P}(X_1)\mathbb{P}(X_2|X_1)\ldots\mathbb{P}(X_n|X_1,\ldots,X_{n-1}) \tag{147}$$

et donc

$$-\frac{1}{n}\log\mathbb{P}(X_1,\ldots,X_n) = -\frac{1}{n}\sum_{i=1}^{n}\log\mathbb{P}(X_i|X_1,\ldots,X_{i-1}) \tag{148}$$

On sent bien que l'entropie conditionnelle (Def. 3) [67] va apparaître dans le membre de droite.

Imaginons un instant que l'on soit dans le cas où

$$-\frac{1}{n}\log\mathbb{P}(X_1,\ldots,X_n) \xrightarrow[n\to\infty]{prob.} \mathbb{H}(\chi) \tag{149}$$

alors utilisant un raisonnement développé par exemple dans le cas continu (Sec. 5.2) on peut définir l'ensemble typique T_n^ε dont le volume/la taille est de l'ordre de $V(T_n^\varepsilon) \approx 2^{n\mathbb{H}(\chi)}$ et la probabilité qu'une donnée appartienne à cet ensemble est quasi-uniforme de valeur $p(\{X\}) \approx 1/V(T_n^\varepsilon)$. Ainsi dans ce cas on a bien un phénomène de concentration des donnés dans un ensemble typique avec une probabilité quasi-uniforme. Ce qui a des conséquences immédiates par exemple en codage, car alors le nombre de bits moyen nécessaires par symbole est de l'ordre de $\mathbb{H}(\chi)$ et cela quand bien même il y a de la dépendance.

Donc, ***l'apparition du phénomène de concentration et ses conséquences est bien plus générale que celle du cas d'indépendance des v.a, mais encore faut-il assurer la convergence quand*** $n \to \infty$**.** Il faut se représenter cette convergence asymptotique par exemple comme si on avait 1 seule image mais dont le nombre de pixels tend vers l'infini.

6.2 Ergodicité, théorème de Birkhoff

Un temps nous allons revenir à quelques notions de base pour bien introduire la notion d'ergodicité (souvent nommée hypothèse ergodique) introduite par L. Boltzmann en 1871 pour sa théorie cinétique des gaz. On définit un espace de probabilités par un triplet $(\Omega, \mathscr{B}, \mathbb{P})$ avec Ω l'univers, $\mathscr{B}$ l'ensemble des ensembles mesurables (tribu) relativement à la mesure P de probabilité. Une variable aléatoire X est alors une fonction (mesurable)

67. ou son pendant différentielle Sec. 5.5.

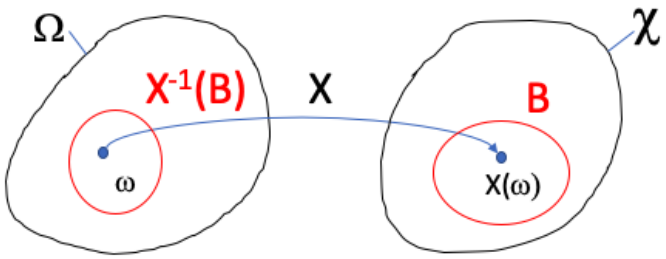


FIGURE 16 – Schéma de la définition d'une variable aléatoire.

de Ω vers χ un espace mesurable. La probabilité associé à X, notée $\mathbb{P}_X$ est définie de telle façon que la probabilité que la v.a ait des valeurs dans $B \subset \chi$ soit la probabilité de l'ensemble $X^{-1}(B)$ (Fig. 16)

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega, x = X(\omega) \in B\}) \tag{150}$$

A ce cadre, on ajoute l'***hypothèse de stationnarité*** (Déf. 7), c'est-à-dire que si l'on prend une collection de $v.a$ alors elles sont toutes de même loi, et les probabilités jointes sont invariantes par translation. Ici cela se traduit par: soit une transformation mesurable T définie sur Ω, alors

$$\forall A \in \mathscr{B},\ \mathbb{P}(T^{-1}(A)) = \mathbb{P}(A) \qquad (stationnarité) \tag{151}$$

(nb. on peut penser à T comme une translation dans le temps). Maintenant à cela, l'hypothèse d'***ergodicité*** ajoute

$$\forall A \in \mathscr{B}\ tq.\ T^{-1}(A) = A \Leftrightarrow \begin{cases} \text{soit } \mathbb{P}(A) & = 0 \\ \text{soit } \mathbb{P}(A) & = 1 \end{cases} \qquad (ergodicité) \tag{152}$$

Cette hypothèse d'ergodicité nous dit que si on applique la transformation T en itérant, en fait on balaye tout l'ensemble, et on ne peut pas être piégé dans une sous-partie de l'ensemble. En effet, mettons que l'on ait un système ayant 2 sous-systèmes invariants par T, l'hypothèse ergodique impose que l'un des deux sous-systèmes soit de mesure nulle, l'on ne peut quasiment pas être piégé dans celui-ci, et l'on vit dans l'autre sous-système.

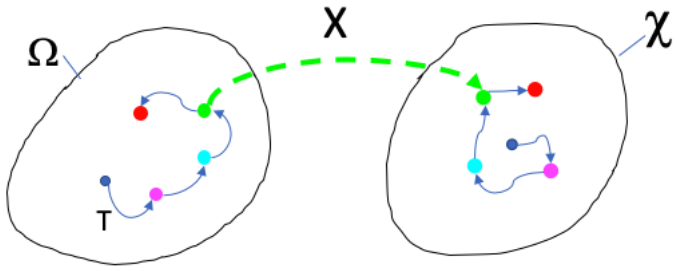


FIGURE 17 – Schéma du processus de moyenne ergodique.

En un sens dans ce sous-système de mesure 1 on a une *notion de mélange*. Notons que si l'on a une transformation qui ne préserve pas la mesure alors ces notions n'ont pas lieu.

Les deux propriétés de stationnarité et d'ergodicité donnent le théorème suivant [68]

Théorème 11 ***(Birkhoff)***

$$\frac{1}{n}\sum_{i=1}^{n} X(T^i(\omega)) \xrightarrow[n\to\infty]{p.s} \mathbb{E}[X] = \int_{\Omega} X(\omega) d\mathbb{P}(\omega) \tag{153}$$

où p.s (presque surement) signifie avec une probabilité 1 [a].

a. $\mathbb{P}(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$

On comprend que lorsque l'on itère T on parcourt l'espace des points dans l'espace Ω qui par action de X parcourt l'espace χ (Fig. 17), l'hypothèse d'ergodicité nous dit alors que la moyenne empirique des valeurs de X calculée sur les points transformés converge vers l'espérance de X relative à la mesure $\mathbb{P}$.

Ainsi si l'on pose la définition suivante

Définition 8 *Soit une transformation ergodique T, on dit que les v.a $(X_1, \ldots, X_n)$ telles que $X_i = X \circ T^i$ forme un processus ergodique.*

68. George David Birkhoff (1884-1944)

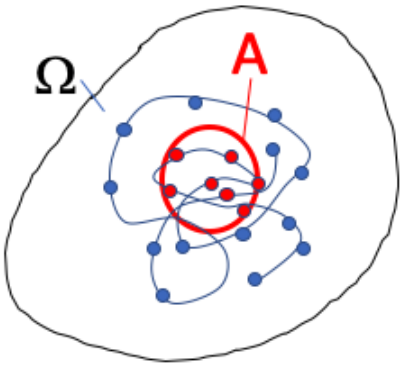


FIGURE 18 – Calcul du temps moyen de séjour dans $A \subset \Omega$.

alors le Th. de Birkhoff nous donne simplement la convergence $p.s$ de la moyenne empirique

$$(X_1, \ldots, X_n)\ ergodiques \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{p.s} \mathbb{E}[X] \tag{154}$$

Un cas particulier est celui du ***temps de séjour*** dans un ensemble $A \subset \Omega$. Pour cela on note $\mathbf{1}_A$ l'indicateur de A, et regardons ce que donne la moyenne des valeurs $\mathbf{1}_A(T^k(\omega))$ (Fig. 18). Le théorème de Birkhoff nous donne alors ($\mathbb{P}(\Omega) = 1$ ici)

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_A(T^k(\omega)) \xrightarrow[n\to\infty]{p.s} \mathbb{E}(\mathbf{1}_A) = \int_\Omega \mathbf{1}_A(\omega) d\mathbb{P}(\omega) = \mathbb{P}(A) \qquad (temps\ de\ séjour) \tag{155}$$

Donc, le nombre de retours moyen dans l'ensemble A par application de la transformation ergodique T, nous donne la probabilité de se trouver dans A, c'est-à-dire sa mesure relativement à la distribution de probabilité, soit en quelque sorte sa taille par rapport à cette mesure.

Cette propriété fondamentale de convergence des sommes vers des espérances se retrouve dans des processus physiques.

6.3 Théorème de Shannon–McMillan–Breiman

Pour en revenir à notre problème initial de convergence Eq. 149, si l'on suppose que les variables $(X_1, \ldots, X_n)$ forment un *processus ergodique* (Déf. 8), certes le logarithme de la probabilité conditionnelle $\mathbb{P}(X_i|X_1 \ldots, X_{i-1})$ est une *v.a* mais on a un problème de bord à cause du X_i qui empêche d'opérer une translation sur les indices (cf. hyp. stationnarité). Si X_i ne dépendait que de $X_{i-\ell} \ldots, X_{i-1}$ $(i > \ell)$ on pourrait tenter une translation, et avec l'aide de l'ergodicité on pourrait alors avoir la convergence *p.s*, donc la convergence en probabilité. Voici le théorème qui permet d'assurer l'apparition des phénomènes de concentrations sur des ensembles typiques:

Théorème 12 *(**Shannon–McMillan–Breiman**) Soit $(X_1, \ldots, X_n)$ un processus stationnaire et ergodique de probabilité $\mathbb{P}$*

$$-\frac{1}{n}\log \mathbb{P}(X_1, \ldots, X_n) \xrightarrow[n\to\infty]{p.s} \mathbb{H}(\chi) \tag{156}$$

Démonstration 12. Nous allons donner les grandes étapes de la démonstration. Si l'on revient à la décomposition de la probabilité jointe

$$-\frac{1}{n}\log \mathbb{P}(X_1, \ldots, X_n) = -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{P}(X_i|X_1, \ldots, X_{i-1}) \tag{157}$$

il nous faut étudier le terme générique de la somme du membre de droite. L'entropie conditionnelle associée, à savoir son espérance, c'est-à-dire $(\{X\} = (X_1, \ldots, X_n))$

$$\mathbb{H}(X_i|X_{i-1}, \ldots, X_1) = \mathbb{E}_{\{X\}\sim\mathbb{P}(\{X\})}[-\log \mathbb{P}(X_i|X_{i-1}, \ldots, X_1)] \tag{158}$$

peut être encadrée en se rappelant que l'entropie augmente (diminue) si on diminue (augmente) le conditionnement (Eq. 121 et Sec. 5.5). Ainsi, pour $1 < \ell$

$$\mathbb{E}[-\log \mathbb{P}(X_i|X_{i-\ell}, \ldots, X_1)] \leq \mathbb{H}(X_i|X_{i-1}, \ldots, X_1) \leq \mathbb{E}[-\log \mathbb{P}(X_i|X_{i-1}, \ldots, X_{-\infty})] \tag{159}$$

si l'on étend la loi de probabilité pour des indexes négatifs jusqu'à $-\infty$. La démonstra-

tion [69] est assez technique. Cependant, la philosophie est de jouer sur cet encadrement, afin de démontrer une convergence des bornes inférieure et supérieure. Cela se fait en utilisant l'hypothèse d'ergodicité et le théorème de Birkhoff. Enfin, en démontrant que les deux limites sont égales, cela assure la convergence de l'entropie conditionnelle prise en sandwich. ■

Ce théorème est central. Pour mémoire, il donne accès à toutes les notions de ***concentration, ensemble typique, uniformité de la probabilité*** avec au cœur l'***entropie*** qui fixe les caractéristiques de ces notions. Les conséquences sont immédiates comme par exemple pour le codage. Cependant, ce théorème peut paraître abstrait et pour mieux comprendre sa signification profonde, nous allons l'étudier dans le cadre des chaines de Markov. Cet outil est très important à la fois conceptuel - dans la compréhension des systèmes dynamiques stochastiques, la limite des chaines de Markov donne accès aux équations différentielles stochastiques - et à la fois algorithmique pour réaliser des l'échantillonnages de distributions de probabilités en grande dimension (eg. Monte Carlo Markov Chain). Donc, nous allons voir un panorama des chaines de Markov qui vont nous servir pour accéder au 2nd Principe de la Thermodynamique avec la croissance de l'entropie.

6.4 Chaines de Markov [70]

6.4.1 Définitions et propriétés

L'idée sous-jacente directrice est l'hypothèse que le futur est accessible en connaissant uniquement le présent (cf. pas de mémoire). Soit alors la définition suivante:

Définition 9 *(**Chaine de Markov**)*

69. Voir Sec. 16.8 du livre de Th. Cover et J. Thomas donné en avant-propos de cette année.
70. Andreï Andreïevitch Markov (1856-1922)

Un processus $(X_1, \dots, X_n)$ est une chaine de Markov si

$$\forall (x_i)_{i \leq n+1} \in \chi^{n+1} \quad \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_1 = x_1) = \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n) \quad (160)$$

Une petite remarque pratique: si on a un passé de taille finie (ex. ℓ) sachant que X_n est la valeur présente, alors on peut définir un processus $Y_n = (X_n, \dots, X_{n-\ell})$ qui sera markovien. Une conséquence immédiate de cette définition est (nb. on utilise par la suite la notation avec les (x_i) que l'on nomme souvent "état" à l'instant/l'étape "i")

$$\begin{aligned} \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = p(x_1, \dots, x_n) &= p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}, \dots, x_1) \\ &= \underbrace{p(x_1)}_{loi\ init.} \prod_{i=2}^{n} \underbrace{p(x_i | x_{i-1})}_{prob.\ transition} \end{aligned} \quad (161)$$

qui fait apparaitre la notion de ***probabilité de transition*** entre deux états voisins. L'hypothèse de stationnarité introduit la notion de ***chaine stationnaire ou homogène***:

Définition 10 *(**Chaine de Markov stationnaire**)*
Une chaine de Markov est stationnaire si

$$\forall (x, y) \in \chi^2, \forall n, \quad \mathbb{P}(X_{n+1} = y | X_n = x) = \mathbb{P}(X_n = y | X_{n-1} = x) \quad (162)$$

donc les probabilités de transition ne dépendent pas de l'instant n, que l'on peut noter $P_{x,y}$ (attention à l'ordre, pensez "x influe sur y").

Propriété 3
Dans le cas d'un alphabet (ie. χ est discret) alors $(P_{x,y})_{(x,y) \in \chi^2}$ est une matrice stochastique de transition qui a les propriétés suivantes:

- *primo*

$$\forall x \in \chi,\ \sum_{y\in\chi} P_{x,y} = 1 \qquad (matrice\ stochastique) \tag{163}$$

qui vient du fait que $p(y|x) = p(x,y)/p(x)$ *et* $\sum_y p(x,y) = p(x)$.

- *secundo, en notant que*[a]

$$\forall y \in \chi,\ \mathbb{P}(X_{n+1} = y) = \sum_{x\in\chi} \mathbb{P}(X_n = x) P_{x,y} \tag{164}$$

on peut utiliser une forme matricielle en regroupant sous forme d'un vecteur colonne μ_n *l'ensemble des probabilités à l'étape* n*, ainsi*

$$\mu_n := (\mathbb{P}(X_n = x))_{x\in\chi} \Longrightarrow \mu_{n+1} = P^T\ \mu_n \tag{165}$$

a. on le voit facilement en exprimant que $\mathbb{P}(X_{n+1} = y)$ est une marginale de la probabilité jointe de (X_{n+1}, X_n).

Quand on fait appelle à l'hypothèse ergodique, le théorème de Birkhoff (Th. 11) qui assure la convergence, nous invite à étudier les itérées de la transformation $P = (P_{x,y})$ qui par exemple donne

$$\mu_{n+k} = (P^T)^k\ \mu_n \tag{166}$$

L'étude des ***valeurs propres de cette matrice*** de transition vont nous renseigner sur les notions d'***équilibre*** et de ***mesure invariante***.

6.4.2 Quelques exemples

6.4.2.1 Modèle à 2 états

Un premier exemple simple (Fig. 19) de processus markovien est celui dont la matrice de transition est donnée par

$$P_{x,y} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \tag{167}$$

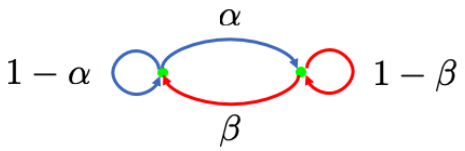


FIGURE 19 – Processus de Markov stationnaire à 2 états.

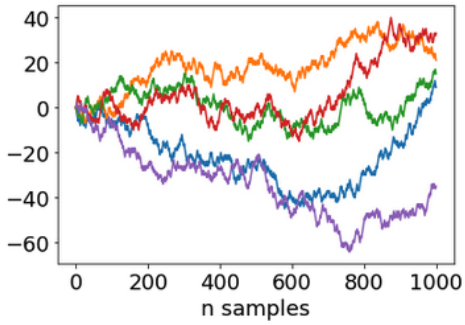


FIGURE 20 – Marche aléatoire 1D: $X_n = X_{n-1} + Z_n$ où Z_n suit une loi de Bernoulli $\{+1, -1\}$ ($p(+1) = 1/2$). Les cinq chaines partent toutes de l'origine $x_1 = 0$. Voir le notebook *randomwalk.ipynb*.

6.4.2.2 Marche aléatoire

On peut étudier des choses plus complexes en utilisant des ***marches aléatoires généralisées*** qui introduisent la notion de variables cachées indépendantes

Théorème 13 *Soit $\{Z_i\}_{i\geq 1}$ des v.a iid et indépendantes des $(X_j)_{j\geq 1}$, l'état à l'étape n dépend d'une fonction récurrente telle que*

$$X_{n+1} = f(X_n, Z_{n+1}) \tag{168}$$

C'est typique des équations différentielles stochastiques où Z_n est une sorte de bruit qui induit des transitions aléatoires. Alors $(X_1, \ldots, X_n)$ est une chaine de Markov stationnaire.

Démonstration 13. Il nous faut montrer l'absence de mémoire, or

$$\mathbb{P}(X_{n+1} = x_{n+1}|X_n = x_n, \dots, X_1 = x_1) = \mathbb{P}(f(x_n, Z_{n+1}) = x_{n+1}|X_n = x_n, \dots, X_1 = x_1) \tag{169}$$

or $(X_n, \dots, X_1)$ dépend de $(Z_n, \dots, Z_1, X_1)$, or Z_{n+1} est indépendant de ces variables donc

$$\begin{aligned}\mathbb{P}(X_{n+1} = x_{n+1}|X_n = x_n, \dots, X_1 = x_1) &= \underbrace{\mathbb{P}(f(X_n = x_n, Z_{n+1}) = x_{n+1})}_{indep.\ des\ états\ 1,\dots,n-1} \\ &= \mathbb{P}(X_{n+1} = x_{n+1}|X_n = x_n)\end{aligned} \tag{170}$$

Donc on a bien un processus markovien. Pour que le processus soit de plus stationnaire il faudrait que $\forall(x, y) \in \chi^2$

$$\mathbb{P}(X_{n+1} = y|X_n = x) = \mathbb{P}(X_n = y|X_{n-1} = x) \Leftrightarrow \mathbb{P}(f(x, Z_{n+1}) = y) = \mathbb{P}(f(x, Z_n) = y) \tag{171}$$

Or,

$$\mathbb{P}(f(x, Z_n) = y) = \int \mathbf{1}_{f(x,z)=y}\, d\mathbb{P}_{Z_n}(z) \tag{172}$$

mais comme les $(Z_i)_i$ sont de même loi, alors le membre de droite est indépendant de n. Ainsi, on a bien une chaine de Markov stationnaire. ■

Un exemple de marche aléatoire sur $\mathbb{Z}$ avec les Z_n des variables de Rademacher $\{+1, -1\}$ $(p(+1) = 1/2)$ est donné dans la figure 20:

$$X_{n+1} = X_n + Z_{n+1} \tag{173}$$

On peut imaginer un automate stochastique comme une machine de Turing qui à chaque fois qu'il lit une nouvelle lettre (instruction) il change d'état et la transition dépend d'une variable aléatoire (inconnue/cachée).

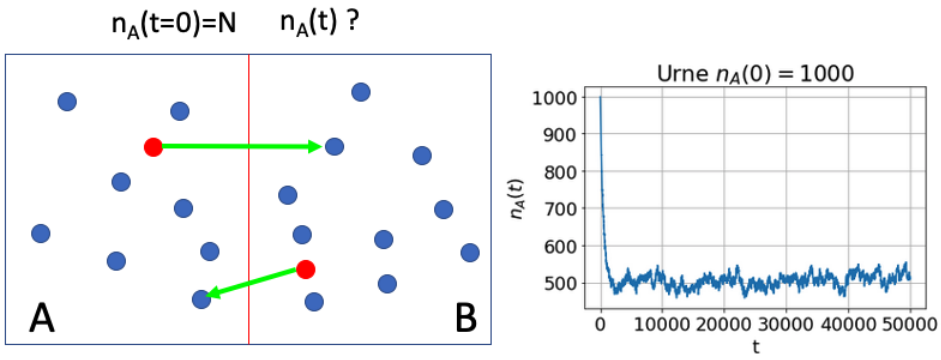


FIGURE 21 – Processus de Markov baptisé "urne d'Erhenfest" comme modèle d'un gaz parfait diffusant à travers une parois poreuse. A chaque étape, on tire une boule au hasard et on la fait changer de compartiment $A \to B$ ou $B \to A$. A droite le nombre de boules qui se trouvent dans l'urne A au bout d'un certain t. (notebook *urne_Ehrenfest.ipynb*)

6.4.2.3 Urne d'Erhenfest

Un autre modèle plus proche de la Physique, est celui de l'Urne d'Erhenfest [71](1907), introduit par Afanasyeva et Paul, qui modélise la diffusion d'un gaz parfait à travers une paroi poreuse. Il est illustré sur la figure 21. Si à l'instant initial, toutes les boules sont dans l'urne A ($n_A(t=0) = N$), qu'advient-il après un temps t sachant qu'à chaque étape discrète du temps on choisit au hasard une boule et la change d'urne: si elle était dans l'urne A on la met dans l'urne B et vice-versa. Le graphe de droite (notebook *urne_Ehrenfest.ipynb*) nous conforte dans l'expérience de tous les jours où l'on observe une homogénéisation à parts égales des distributions dans les 2 urnes. ***Il y a un phénomène d'irréversibilité qui s'installe***. Attention, cela ne veut pas dire qu'une boule ne peut pas franchir la paroi, mais en moyenne il y en a autant qui font le chemin de $A \to B$ et $B \to A$ après un "certain temps". ***Mais ce comportement d'irréversibilité quand N est petit n'apparait pas, comme illustré sur la figure 22; il y a de fortes chances alors de se retrouver dans l'état initial! ce qui est extrêmement peu probable quand $N \gg 1$.*** On peut en effet montrer que le temps moyen entre 2 retours à l'état initial est $\langle \tau \rangle = 2^N$.

71. Paul Ehrenfest (1880-1933) eu d'importantes contributions notamment en Mécanique Quantique et Relativité avec sa femme Afanasyeva et sa fille Tatyana.

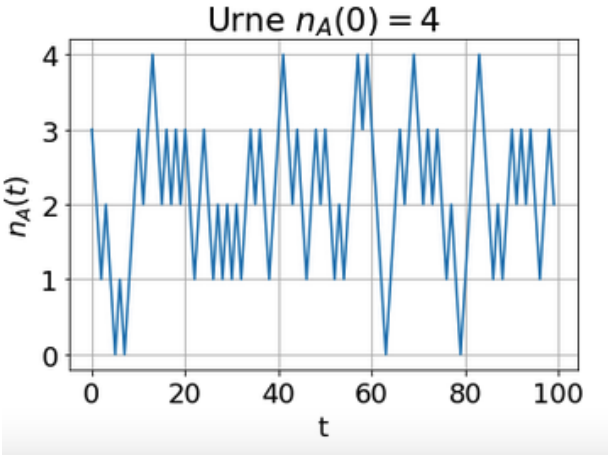


FIGURE 22 – Même expérience que la figure 21 mais avec $N = 4$, dans ce cas le retour à l'état initial n'est pas du tout négligeable.

La modélisation se fait aisément si $X_n = n_A(t = n)$, et posant $X_n = x$ alors

$$X_{n+1} = \begin{cases} x-1 & proba.\ \frac{x}{N} \\ x+1 & proba.\ 1-\frac{x}{N} \end{cases} = X_n + Z_{n+1} \tag{174}$$

où $Z_{n+1} \in \{-1, +1\}$ mais ce n'est pas une variable de Rademacher car

$$\mathbb{P}(Z_{n+1} = -1 | X_n = x) = \frac{x}{N} \tag{175}$$

donc il n'y a pas indépendance de la loi Z_{n+1} vis-à-vis de l'état X_n. Ainsi tout en étant une chaine de Markov, le processus n'est pas une marche aléatoire définie dans le précédent paragraphe. Il est intéressant de voir alors ***dans quel cas apparait un équilibre et le phénomène d'irréversibilité***.

6.5 Loi invariante ou stationnaire: équilibre

Dans le cadre des processus markoviens, à la lueur par exemple de l'urne d'Erhenfest, les distributions de probabilités évoluant dans le temps (n), on peut se demander s'il y a au bout d'un certain temps $(n \geq n_o)$ une forme asymptotique qui rendrait le système

invariant. Soit la définition suivante:

Définition 11 *(**loi invariante**)*
Une loi invariante, notée Π*, est une mesure de probabilité telle que*[a] *(Prop. 3, Eqs. 163, 164) (P matrice de transition* $(P_{x,y})_{(x,y)\in\chi^2}$*)*

$$\sum_{x\in\chi}\Pi(x)=1 \qquad\qquad \Pi(y)=\sum_{x\in\chi}\Pi(x)P_{x,y} \tag{176}$$

En d'autres termes

$$\Pi = P^T\Pi \tag{177}$$

c'est-à-dire que Π *est un invariant par action de la matrice de transition, c'est un vecteur propre associé à la valeur propre* $\lambda = 1$*.*

a. nb. c'est comme si dans l'équation 164 on faisait $n \to \infty$.

Par exemple, dans le processus à 2 états ($\{x, y\}$) (Sec. 6.4.2.1), pour mémoire

$$P_{x,y} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \tag{178}$$

dont la transposée a 2 valeurs propres $(1, 1-\alpha-\beta)$. Donc, on a bien une distribution invariante ($\lambda = 1$) satisfaisant non seulement la condition d'unitarité mais aussi

$$-\alpha\Pi(x)+\beta\Pi(y)=0 \tag{179}$$

donc

$$\Pi(x)=\frac{\beta}{\alpha+\beta} \qquad\qquad \Pi(y)=\frac{\alpha}{\alpha+\beta} \tag{180}$$

D'une manière générale on a $|\chi|$ équations (la condition d'unitarité est redondante car P en satisfait une également) à $|\chi|$ inconnues, donc à moins de se retrouver dans un cas dégénéré, il existe une seule distribution invariante.

Dans le cas de l'urne d'Erhenfest (Sec. 6.4.2.3) si x est le nombre de boules dans la

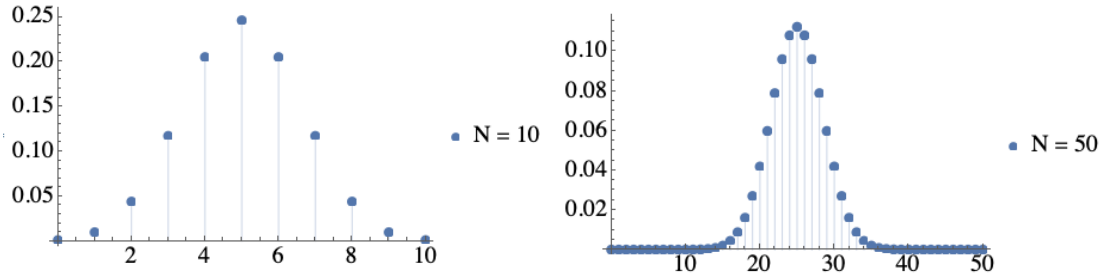


FIGURE 23 – Distribution invariante dans le cas où l'urne d'Erhenfest contient $N = 10$ ou $N = 50$ boules.

boite A, alors l'équation qui régit $\Pi(x)$ est donnée par

$$\Pi(x) = \Pi(x-1)P_{x-1,x} + \Pi(x+1)P_{x+1,x} \tag{181}$$

$$= \Pi(x-1)\left(1 - \frac{x-1}{N}\right) + \Pi(x+1)\frac{x+1}{N} \tag{182}$$

Si l'on fixe la condition $\Pi(0) = \Pi(1) \times \frac{1}{N}$, alors on peut montrer que

$$\Pi(x) = \binom{N}{x}\left(\frac{1}{2}\right)^N \tag{183}$$

C'est une loi Binomiale de probabilité $p = 1/2$. Ceci nous conforte dans l'intuition qu'à l'équilibre les boules se répartissent équitablement entre les deux boîtes. Voir une illustration de cette loi Binomiale pour $N = 10$ et $N = 50$ sur la figure 23. Quand N est grand la loi converge vers une loi normale $\mathcal{N}(\mu = N/2; \sigma^2 = N/4)$.

La mesure de probabilité Π ***est stable pour le système, toute transformation via*** P ***la laisse inchangée***. Que se passe-t-il quand on part d'une situation déséquilibrée? Dans les simulations numériques du paragraphe précédent (Sec. 6.4.2.3) on a remarqué que le système peut évoluer vers cette situation stable en moyenne mais qu'il y a des fluctuations particulièrement sévères quand N est petit. Mais, quand N est grand, on observe une sorte d'*irréversibilité*, c'est-à-dire qu'il y a très peu (voire pas en pratique) de possibilité de se retrouver dans l'état où toutes les boules sont dans le même compartiment. Donc, plus généralement, on s'intéresse à ***la dynamique du système pour savoir quand et comment il peut converger vers une situation stable***. La notion de réversibilité des lois

tout en observant un phénomène d'irréversibilité a été assez paradoxale dans les premiers temps de la Thermodynamique. Et la notion d'entropie (Carnot) est apparue en lien avec les observations de systèmes physiques. Voyons ce qu'il en est d'un point de vue mathématique.

6.6 Loi stationnaire/invariante et la réversibilité

L'idée sous-jacente est que l'existence de la loi invariante dépend de la réversibilité. Mais comment définir la réversibilité d'une chaine de Markov? Pour une chaine de Markov, on définit la transition de X_n à X_{n+1}, donc inverser le temps nécessite de définir comment l'on passe de X_{n+1} à X_n.

Définition 12 ***(Matrice inverse)***
Soit un processus stationnaire (Déf. 10) avec une matrice $P_{x,y} = \mathbb{P}(X_{n+1} = y | X_n = x)$ $(\forall (x,y) \in \chi^2)$*. S'il existe une loi stationnaire* $\Pi(x) \neq 0$*, c'est-à-dire que tous les* x *sont accessibles alors on définit la matrice* $Q_{x,y}$ *suivante*

$$Q_{x,y} = \frac{\Pi(y) P_{y,x}}{\Pi(x)} \tag{184}$$

Alors Q est une **matrice stochastique** (Eq. 163). En effet, comme $\Pi = P^T \Pi$, alors $\Pi_x = \sum_y \Pi_y P_{y,x}$ donc $\sum_y Q_{x,y} = 1$. De plus Q est la matrice de transition de X_{n+1} à X_n pour la distribution invariante, en effet via la formule de Bayes

$$\begin{aligned} \mathbb{P}(X_n = y | X_{n+1} = x) &= \frac{\mathbb{P}(X_{n+1} = x | X_n = y) \mathbb{P}(X_n = y)}{\mathbb{P}(X_{n+1} = x)} \\ &= P_{y,x} \frac{\mathbb{P}(X_n = y)}{\mathbb{P}(X_{n+1} = x)} \end{aligned} \tag{185}$$

Si maintenant, on a initialisé le processus par la distribution invariante Π, alors

$$\mathbb{P}(X_n = y | X_{n+1} = x) = P_{y,x} \frac{\Pi(y)}{\Pi(x)} = Q_{x,y} \tag{186}$$

Ainsi, si la chaine de Markov est générée dans les conditions où $X_1 \sim \Pi$ et on applique P successivement pour donner $X_2, \ldots, X_n$, toutes les variables X_i suivent la loi de distri-

bution Π, et en retour on peut revenir en arrière de X_n à $X_{n-1}, \ldots, X_1$ en appliquant Q. Ce qui se résume par ce petit schéma

$$X_1(\Pi) \underset{Q}{\overset{P}{\rightleftharpoons}} X_N(\Pi)$$

Notons que la loi Π est telle que

$$\Pi = P^T\Pi = Q^T\Pi \tag{187}$$

mais en tout état de cause il n'y a pas lieu de conclure que $P = Q$. D'où la définition suivante qui va spécifier le cadre qui nous intéresse de lois de la Physique réversibles (sauf l'interaction faible par exemple [72])

Définition 13 *(**Chaine de Markov réversible**)*
On dit qu'une chaine de Markov stationnaire de matrice de transition P est réversible relativement à une probabilité invariante Π, si étant dans les conditions de la définition 12 alors $Q = P$, ce qui se traduit par[a]

$$\Pi(x)P_{x,y} = \Pi(y)P_{y,x} \qquad (\textit{balance détaillée}) \tag{188}$$

(nb. la balance globale est donnée par l'équation $\Pi = P^T\Pi$) qui se traduit par le petit schéma

$$X_1(\Pi) = x \underset{P_{y,x}}{\overset{P_{x,y}}{\rightleftharpoons}} X_2(\Pi) = y$$

a. NDJE. J. C. Maxwell en 1867 a introduit cette notion dans son étude de la cinétique des gaz (*principle of sufficient reason*).

La propriété suivante est particulièrement intéressante en pratique

Propriété 4 *(**existence d'une mesure invariante**)*
Si un processus de Markov stationnaire de probabilité de transition P a une distribution Π satisfaisant la balance détaillée, alors Π est une mesure invariante (ie. loi

72. NDJE. Les lois du Modèles Standard conservent CPT avec une très grande précision (C conjugaison de charge, P parité et T renversement du temps, mais depuis les expériences sur les Kaons neutres de 1964, on sait que l'interaction faibles brisent l'invariance CP donc T.

stationnaire).

La démonstration est simple, on sait que $\forall x, y$, $\Pi(x)P_{x,y} = \Pi(y)P_{y,x}$, sommons sur y de part est d'autre

$$\sum_y \Pi(y)P_{y,x} = \sum_y \Pi(x)P_{x,y} = \Pi(x) \underbrace{\sum_y P_{x,y}}_{=1 \ (mat.\ stoch.)} = \Pi(x) \tag{189}$$

qui n'est rien d'autre que $\Pi = P^T\Pi$, donc Π est bien invariante (Prop. 11).

Les marches aléatoires ainsi que l'urne d'Erhenfest sont des exemples de processus inversibles. Dans la séance prochaine nous verrons comment l'hypothèse d'ergodicité permet d'une part que tous les états soient explorés, et que d'autre part la chaine ne reste pas bloquée dans un état récurent. Ainsi, cela permettra d'assurer la convergence vers la mesure invariante. Alors, ayant ces propriétés en main, nous aurons accès au taux d'entropie, à la dynamique de l'entropie et finalement à l'évolution vers un maximum d'entropie.

6.7 NDJE. Matrice stochastique, loi stationnaire et réversibilité

Nous avons vu qu'une matrice stochastique P satisfait par définition $\forall x$, $\sum_y P_{x,y} = 1$ (Eq. 163).

- si $P = P^T$ cela implique nécessairement l'existence d'une loi stationnaire Π uniforme et le processus est réversible. En effet, si $\Pi(y) = C \neq 0$ pour tout y, $\Pi(x) = C = \sum_y P_{y,x}\Pi(y)$ donne $\sum_y P_{x,y} = 1$ ce qui est vrai.
- si l'on dispose d'une matrice stochastique P qui de plus satisfait pour tout x, $\sum_y P_{y,x} = 1$, on dit quelle est *bi-stochastique* et en reprenant la démonstration précédente, on montre que la loi uniforme est stationnaire. La différence est que le processus n'est pas nécessairement réversible [73], car dans ce cas on peut avoir $Q = P^T \neq P$. Un tel cas de figure est par exemple donné par un

73. Un cas bi-stochastique à 3 états où $Q = P^T$ donc réversible est par exemple

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}$$

.

processus régit par les matrices P et Q suivantes

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \qquad Q = P^T = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \neq P \tag{190}$$

que l'on peut représenter par le schéma 24.

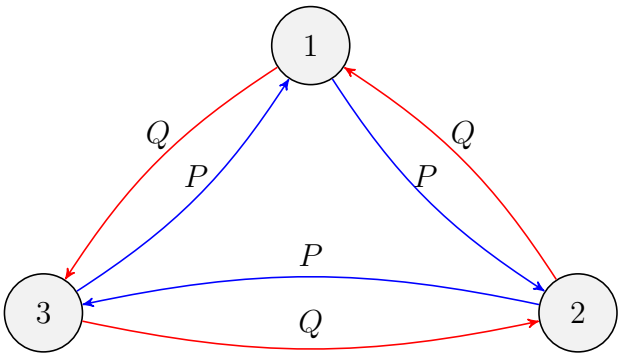


FIGURE 24 – Processus markovien de matrices de transition *forward* P et *backwrd* Q.

7. Séance du 22 Fev.

Nous allons étudier la dynamique d'un système, ici sous forme d'une ***chaine de Markov***, qui tend vers l'équilibre où l'entropie atteint son maximum. Rappelons que dans ce cadre (Def. 9), ***le futur du système n'est conditionné que par son état présent***, c'est-à-dire qu'il ne dépend pas de ses états passés. Typiquement, nous allons trouvé ce type de comportement lors de la discrétisation d'équations différentielles où n'intervient que la dérivée première par rapport au temps. De plus, nous considérerons les ***chaines stationnaires*** (Déf. 10), où ***la probabilité de transition*** X_n ***vers*** X_{n+1} ***ne dépend pas du "temps"***. On a alors une grande matrice [74] potentiellement infinie $(P_{x,y})_{(x,y)\in\mathcal{X}^2}$ (notée par la suite P). Pour mémoire, $P_{x,y}$ ***est une matrice stochastique*** satisfaisant une relation de normalisation sur les y (Eq. 163). Cette matrice régit la dynamique de génération de la

74. Rappel mnémotechnique: *"x influe sur y"*.

chaine, et nous avons vu que si l'on note $\mu_n = (\mathbb{P}(X_n = x))_{x\in\chi}$ alors (Eq. 165)

$$\mu_{n+1} = P^T \mu_n$$

A l'équilibre (s'il existe), il y a une ***mesure invariante*** Π qui satisfait l'équation [75] (Déf. 11)

$$\Pi = P^T \Pi$$

C'est un vecteur propre de P^T avec la valeur propre $\lambda = 1$. Cette mesure invariante nous dit que si à l'instant initial $t = 1$ le système, par exemple l'ensemble des vitesses des corpuscules d'un gaz, est tel que $X_1 \sim \Pi$ alors à l'instant $t = n$, $X_n \sim \Pi$ également, ce qui est la définition d'un état d'équilibre.

La notion de ***réversibilité***, nous a amené à considérer la matrice Q (Déf. 12) définie à partir de P et de la mesure invariante Π selon

$$Q_{x,y} = P_{y,x} \frac{\Pi(y)}{\Pi(x)}$$

Elle fait évoluer le système à rebours (inversion du temps). Dans le cas de figure où une distribution de probabilité satisfait la relation dite de ***balance détaillée*** (Déf. 13, Eq. 188)

$$\Pi(x) P_{x,y} = \Pi(y) P_{y,x}$$

alors Π est une mesure invariante. On peut le formuler comme *"il y a autant d'états qui transitent de x vers y que d'états qui transitent de y vers x"*.

7.1 Marche aléatoire sur un graphe non directionnel

Considérons des graphes non-directionnels (Fig. 25) où par définition les transitions $x \to y$ ont le même poids que les transitions $y \to x$. Ces graphes peuvent représenter les probabilités de transitions de chaines de Markov stationnaires. Les éléments de la matrice P sont

$$P_{x,y} = \frac{W_{x,y}}{\sum_z W_{x,z}} \tag{191}$$

75. NDJE. on peut également utiliser la transposée $\Pi^T = \Pi^T P$.

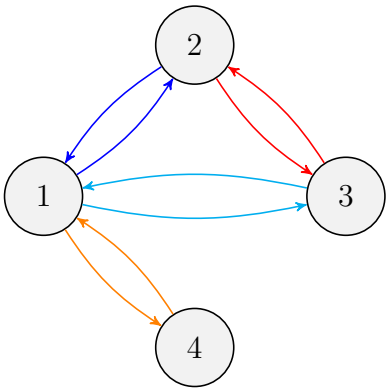
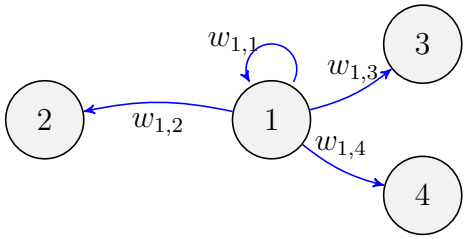


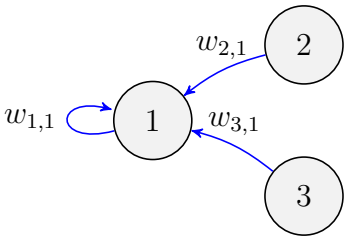
FIGURE 25 – Exemple de graphe non directionnel où pour tout couple (x, y), la transition $x \to y$ a le même poids que la transition $y \to x$.

c'est-à-dire que l'on renormalise par la somme des poids des flèches qui partent de x comme sur le schéma ci-dessous.



Qu'en est-il de la mesure invariante? Faisons le raisonnement suivant: soit la somme des poids de toutes les flèches qui pointent sur x

$$W_x := \sum_y W_{y,x} \tag{192}$$



elle représente une sorte de mesure de l'attraction de x. Normalisons la par $\sum_x W_x = \sum_{x,y} W_{x,y} := W$ pour obtenir une probabilité, et définissons alors

$$\Pi(x) = \frac{W_x}{W} \tag{193}$$

Alors on constate que

$$\Pi(x)P_{x,y} = \frac{\sum_z W_{z,x}}{W} \times \frac{W_{x,y}}{\sum_z W_{x,z}} \tag{194}$$

Or pour tout couple (x, z), $W_{z,x} = W_{x,z}$ donc

$$\Pi(x)P_{x,y} = \frac{W_{x,y}}{W} = \frac{W_{y,x}}{W} = \Pi(y)P_{y,x} \tag{195}$$

Donc, Π satisfait la balance détaillée, c'est donc une mesure invariante.

Ce cas particulier des graphes non-directionnels qui donnent des chaines de Markov réversibles est celui que l'on trouve en Physique, c'est ainsi qu'il nous intéresse au premier chef.

Cependant, on peut se demander si la loi stationnaire (mesure invariante) existe toujours? Nous avons le cas général le théorème de Shannon–McMillan–Breiman (Th. 12) qui nous assure la convergence presque surement de $-\frac{1}{n}\log \mathbb{P}(X_1, \ldots, X_n)$ et donne lieu au phénomène de concentration sur les ensembles typiques, mais voyons ce qu'il en est dans le cadre markovien. Pour cela, l'***ergodicité*** entre en jeu (Birkhoff, Th. 11).

7.2 Ergodicité et chaine de Markov

Rappel, dans le cadre d'un processus ergodique régit par une loi de transformation T, on atteint l'état X_n en appliquant n fois la transformation T à l'état X_1, et d'une manière générale $X_{n+k} = T^k X_n$, ce qui nous assure par le théorème de Birkhoff que (Eq. 154)

$$(X_1, \ldots, X_n)\ ergodiques \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{p.s} \mathbb{E}[X]$$

Maintenant, introduisons les définitions suivantes qui vont nous permettre d'assurer l'existence de la loi stationnaire.

Définition 14 *(**chaine irréductible**)*
Une chaine de Markov est irréductible si on peut passer de n'importe état x à n'importe quel état y, c'est-à-dire

$$\forall (x, y) \in \chi^2, P_{x,y} > 0 \tag{196}$$

Notons qu'il n'y a pas forcément un lien direct entre x et y: par exemple dans le schéma 25, l'état "4" n'a pas de lien direct avec les états "2" et "3", cependant la chaine est irréductible. Si l'équation 155 nous donne le temps de séjour d'une chaine de Markov dans un ensemble A, définissons plutôt ici *le temps de premier retour* à un état quelconque.

Définition 15 *(**temps de premier retour**)*
Pour un état $X_1 = x$, on définit le "temps" T_x

$$T_x = \inf\{n; X_n = x\} \tag{197}$$

si l'on ne peut atteindre x alors $T_x = +\infty$.

NDJE. On peut se faire une idée par exemple: on tire au hasard un état $x_1 \in \chi$ à l'instant $t = 1$ selon la loi de X_1; puis pour déterminer à l'instant $t = 2$ le nouvel état x_2, on tire au hasard selon les probabilités $(P_{x_1,x})_{x\in\chi}$ de la matrice de transition P, et ainsi de suite, on note alors quand est-ce que l'on retrouve l'état initial, notons le $n^{(1)}$. On recommence K fois cet exercice, T_x est alors le temps minimum des $(n^{(k)})_{k\leq K}$, ou plutôt l'infinimum obtenu car on considère $K = +\infty$.

S'il y a un nombre fini d'états et que tous sont accessibles à partir de x_1, alors T_x est fini, mais dans le cas où il y a un nombre infini d'états, la question se pose. Remarquons que T_x est lui-même une *v.a.* En effet, même si l'on part toujours du même état initial, à chaque nouvelle évolution de la chaine de Markov, les états intermédiaires successifs sont tirés aléatoirement selon P avant de revenir en x_1. Autrement dit, il y a plusieurs chemins possibles qui partent de x_1 pour y revenir. Soit alors la définition suivante:

Définition 16 *(état récurent positif)*
Un état x est dit récurent positif si l'espérance de son temps de retour est fini[a]

$$\mathbb{E}_x[T_x] < \infty \tag{198}$$

a. NDJE. en ayant à l'esprit $\mathbb{P}(X_1 = x) = 1$, c'est-à-dire que l'on commence les évolutions des chaines toujours par l'état x.

Cela veut dire que dans le cas d'une chaine irréductible, partant de l'état x, on peut y revenir avec un nombre arbitrairement grand de fois avec une fréquence $1/T_x$ qui mesure l'attractivité de l'état x. Or, suivant x, cette attractivité change, il y a des états qui drainent plus de flèches que d'autres. Dans le cas de la marche aléatoire, c'est la philosophie sous-jacente à la définition de W_x. Il est alors tentant de se demander si la loi stationnaire a un lien avec cette notion de temps de retour.

Théorème 14 *(**Ergodicité**)*
Soit (X_n) une chaine de Markov, irréductible et récurrente positive, alors

1. *considérant la moyenne du nombre de fois où $X_k = x$*

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{X_k=x} \xrightarrow[n\to\infty]{p.s} \Pi(x) > 0 \tag{199}$$

   où $\Pi(x)$ est une mesure invariante;

2. *la mesure invariante précédente est l'unique mesure invariante;*

3. *pour n'importe quelle fonction f, tq. $\sum_x |f(x)|\Pi(x) < \infty$, alors le théorème de Birkhoff nous dit que*

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \xrightarrow[n\to\infty]{p.s} \sum_x f(x)\Pi(x) = \mathbb{E}_{X\sim\Pi}[f(X)] \tag{200}$$

   Si on part de x_1 en $t = 1$ en tirant X_1 selon sa loi, on calcule $f(x)$, ensuite on fait transiter l'état de $t = 1$, à $t = 2$ on calcule $f(x_2)$ et ainsi de suite jusqu'à $f(x_n)$, ensuite on calcule la moyenne empirique des $(f(x_i))_{i\leq n}$. Quand n tend vers l'infini, le résultat est identique à faire la moyenne des $f(x)$, si on tire x selon la loi invariante Π. C'est-à-dire que la moyenne temporelle est identique

> *à la moyenne d'ensemble. En fait, si on rajoute que la chaine est apériodique, alors n'importe quelle mesure de départ converge vers la loi invariante.*

La démonstration de ce théorème est repoussée à un autre cours. Il est important car si l'on a un processus se décrivant par l'évolution d'une chaine de Markov, moyennant les deux hypothèses d'irréductibilité et de récurrence positive, alors il existe un état d'équilibre vers lequel le processus converge (en loi). Ce qui donne accès à l'entropie à l'équilibre.

7.3 Entropie d'une chaine de Markov à l'équilibre

Nous allons revenir sur des notions de ***taux d'entropie*** vues aux sections 5.4.1, 5.4.2 et en particulier le théorème 9 valant pour les processus stationnaires:

$$\mathbb{H}(\chi) = \mathbb{H}'(\chi) = \lim_{n\to\infty} \mathbb{H}(X_n|X_{n-1}, \ldots, X_1) \tag{201}$$

Dans le cas de *chaines de Markov*, le problème se simplifie car l'entropie conditionnelle se réduite à $\mathbb{H}(X_n|X_{n-1})$. De plus, dans *le cas stationnaire*, cette entropie ne dépend pas de n, elle est égale par exemple à $\mathbb{H}(X_2|X_1)$. A l'équilibre, X_2 et X_1 ont même loi, en l'occurrence Π. Donc, selon la définition de l'entropie conditionnelle [76] (Déf. 3), on a

$$\mathbb{H}^{(\prime)}(\chi) = \mathbb{H}(X_2|X_1) = \sum_{x\in\chi} \mathbb{P}(X_1 = x)\mathbb{H}(X_2|X_1 = x) \tag{202}$$

Or, à l'équilibre $\mathbb{P}(X_1 = x) = \Pi(x)$. Maintenant

$$\begin{aligned}\mathbb{H}(X_2|X_1 = x) &= -\sum_{y\in\chi} \mathbb{P}(X_2 = y|X_1 = x) \log \mathbb{P}(X_2 = y|X_1 = x) \\ &= -\sum_{y} P_{x,y} \log P_{x,y}\end{aligned} \tag{203}$$

Il vient alors

$$\mathbb{H}^{(\prime)}(\chi) = -\sum_{(x,y)\in\chi} \Pi(x) P_{x,y} \log P_{x,y} \tag{204}$$

76. NDJE. il peut se faire qu'il y ait des changements de notations en cours de route.

Ainsi, avec une chaine de Markov stationnaire à l'équilibre, on peut donc tout calculer explicitement. La question qui vient est de savoir comment le processus évolue quand on se place hors de l'équilibre? et d'où vient le 2nd Principe de la Thermodynamique?

7.4 Chaine de Markov et 2nd Principe de la Thermodynamique

En préambule, rappelons que si l'on se place dans le cas de processus markoviens, ce cadre est bien adapté pour les processus physiques.

Pour mémoire, le point de vue de Boltzmann (Sec. 4.1) amenait à considérer l'entropie d'un système d'énergie fixée (à l'équilibre) comme le log du nombre de micro-états accessibles, parce que la probabilité est uniforme. Ce que l'on va découvrir, c'est qu'en effet ***quand la distribution de probabilité à l'équilibre est uniforme alors l'entropie augmente vers son maximum***, mais on va également s'apercevoir que ***si la distribution de probabilité à l'équilibre n'est pas uniforme, alors il n'y a pas d'augmentation d'entropie***. Donc attention.

Considérons, à l'instant n deux distributions $\mu_n = (\mathbb{P}(X_n = x))_{x\in\chi}$ et μ'_n. Quelle est l'évolution comparée de ces lois sachant qu'elles sont régies par la même "matrice" de probabilité de transition $(P_{x,y})_{(x,y)\in\chi^2}$? Donc à l'instant n, on dispose de (μ_n, μ'_n) et à l'instant $n+1$ nous avons (μ_{n+1}, μ'_{n+1}), il est alors intéressant de comparer l'évolution de la divergence de Kullback-Leibler entre ces lois (Déf. 4). On sait que la distance est positive ou nulle. Ce que l'on va montrer c'est que la distance décroit en fonction de n.

Théorème 15 *(**rapprochement des lois**)*
Soit $\mu_n \xrightarrow{P} \mu_{n+1}$ *et* $\mu'_n \xrightarrow{P} \mu'_{n+1}$ *alors*

$$D_{KL}(\mu_n \| \mu'_n) \geq D_{KL}(\mu_{n+1} \| \mu'_{n+1}) \tag{205}$$

Démonstration 15.
Si l'on considère deux v.a X_n et X_{n+1}, on peut évaluer leur probabilité jointe, notée ici [77] $p(X_n, X_{n+1})$ dans le cas d'une évolution de μ, et $q(X_n, X_{n+1})$ dans le cas de μ'. Qu'en

77. NDJE. on essaye d'alléger les notations...

est-il de leur divergence?

$$
\begin{aligned}
D_{KL}(p(X_n, X_{n+1}) \| q(X_n, X_{n+1})) &= \sum_{x,y} p(X_n = x, X_{n+1} = y) \log \frac{p(X_n = x, X_{n+1} = y)}{q(X_n = x, X_{n+1} = y)} \\
&= \sum_{x,y} p(X_n = x) p(X_{n+1} = y | X_n = x) \log \frac{p(X_n = x) p(X_{n+1} = y | X_n = x)}{q(X_n = x) q(X_{n+1} = y | X_n = x)} \\
&= \sum_{x,y} p(X_n = x) p(X_{n+1} = y | X_n = x) \log \frac{p(X_n = x)}{q(X_n = x)} \\
&\quad + \sum_{x,y} p(X_n = x) p(X_{n+1} = y | X_n = x) \log \frac{p(X_{n+1} = y | X_n = x)}{q(X_{n+1} = y | X_n = x)} \\
&= \sum_{x} p(X_n = x) \log \frac{p(X_n = x)}{q(X_n = x)} \overbrace{\sum_{y} p(X_{n+1} = y | X_n = x)}^{=1} \\
&\quad + \sum_{x} p(X_n = x) \sum_{y} p(X_{n+1} = y | X_n = x) \log \frac{p(X_{n+1} = y | X_n = x)}{q(X_{n+1} = y | X_n = x)}
\end{aligned}
$$

Le premier terme n'est autre que $D_{KL}(p(X_n) \| q(X_n))$. Concernant le second terme, $p(X_{n+1} = y | X_n = x) = q(X_{n+1} = y | X_n = x) = P_{x,y}$, car c'est la même matrice de transitions (sens *forward*) par hypothèse qui régit l'évolution des deux chaines, donc le second terme de droite est nul. Il vient

$$
D_{KL}(p(X_n, X_{n+1}) \| q(X_n, X_{n+1})) = D_{KL}(p(X_n) \| q(X_n)) \tag{206}
$$

On aurait tout aussi bien pu écrire

$$
\begin{aligned}
&D_{KL}(p(X_n, X_{n+1}) \| q(X_n, X_{n+1})) \\
&\quad = \sum_{xy} p(X_{n+1} = x) p(X_n = y | X_{n+1} = x) \log \frac{p(X_{n+1} = x) p(X_n = y | X_{n+1} = x)}{q(X_{n+1} = x) q(X_n = y | X_{n+1} = x)} \\
&\quad = D_{KL}(p(X_{n+1}) \| q(X_{n+1})) \\
&\qquad + \sum_{x} p(X_{n+1} = x) \sum_{y} p(X_n = y | X_{n+1} = x) \log \frac{p(X_n = y | X_{n+1} = x)}{q(X_n = y | X_{n+1} = x)}
\end{aligned}
$$

Or, $p(X_n = y | X_{n+1} = x)$ et $q(X_n = y | X_{n+1} = x)$ sont régis par des matrices de transitions *backward* qui ne sont par forcément les mêmes (nous ne sommes pas à l'équilibre), donc le second membre n'est pas nul *a priori*. Cependant, il s'écrit comme l'espérance de

$D_{KL}(p(X_n = y|X_{n+1} = x)\|q(X_n = y|X_{n+1} = x))$, donc ce terme est positif ou nul. Ainsi,

$$D_{KL}(p(X_n)\|q(X_n)) \geq D_{KL}(p(X_{n+1})\|q(X_{n+1})) \tag{207}$$

■

Donc l'évolution des chaines ayant les mêmes probabilités de transitions forward, rapproche au cours du temps leurs distributions de probabilités.

Un cas particulier du théorème est celui où ***l'une des deux lois est la mesure invariante***, ex. $\forall n, \mu'_n = \Pi$, alors

$$\forall n, \quad 0 \leq D_{KL}(\mu_{n+1}\|\Pi) \leq D_{KL}(\mu_n\|\Pi) \tag{208}$$

Un cas encore plus particulier est celui où ***la mesure invariante est la loi uniforme***, c'est-à-dire $\Pi(x)$ est indépendante de $x \in \chi$, elle vaut $1/|\chi|$, alors

$$D_{KL}(\mu_n\|\Pi) = \sum_x p(X_n = x) \log \frac{p(X_n = x)}{\Pi(x)} = -\mathbb{H}[\mu_n] - \log|\chi| \tag{209}$$

et donc

Théorème 16 *(**croissance de l'entropie**)*
*Soit $\mu_n \xrightarrow{P} \mu_{n+1}$, s'il existe une **mesure invariante uniforme** alors*

$$\mathbb{H}[\mu_{n+1}] \geq \mathbb{H}[\mu_n] \tag{210}$$

Dans le cas où Π est certes ***une mesure invariante mais n'est pas uniforme***, alors la suite des divergences $D_{KL}(\mu_n\|\Pi)$ va converger (suite positive décroissante) mais ***l'entropie de μ_∞ n'est pas maximale***. Ainsi, on peut se poser la question: à quelle condition la loi invariante est uniforme? Pour ce faire, il nous faut donner des caractéristiques supplémentaires au processus markovien.

Définition 17 *(**matrice bi-stochastique**)*
Soit $P = (P_{x,y})_{(x,y)\in\chi^2}$, elle est doublement stochastique ssi

$$\sum_y P_{x,y} = 1 \quad \text{et} \quad \sum_x P_{x,y} = 1 \tag{211}$$

Ainsi nous avons les propriétés suivantes:

Propriété 5

1) Si la mesure Π invariante et uniforme, alors la P est doublement stochastique.

2) Si P est doublement stochastique, alors la mesure uniforme est invariante.

En effet, pour 1) P est stochastique par définition, et de plus

$$\Pi = P^T\Pi \Rightarrow \Pi(y) = \frac{1}{|\chi|} = \sum_x \Pi(x)P_{x,y} = \frac{1}{|\chi|}\sum_x P_{x,y} \tag{212}$$

donc $\sum_x P_{x,y} = 1$. Pour 2), on reprend juste l'expression précédente pour conclure immédiatement.

Nous avons donc une idée du cadre mathématique qui consolide l'intuition de la Physique. Celle-ce nous dit que l'entropie augmente au fur et à mesure que le temps avance. Si le système est de type markovien, alors à chaque étape de la progression temporelle, conditionné par l'état n, l'état $n+1$ est plus incertain. Les lois se rapprochent de la loi invariante qui diffuse les probabilités sur les micros états disponibles. L'entropie atteinte est maximale si la loi invariante est uniforme (c'est-à-dire la diffusion est maximale). Et l'outil qui nous a servi est l'entropie relative.

Maintenant, nous allons revenir à la modélisation des données en grande dimension, avec l'idée que si ces données se concentrent sur des surfaces, il y a sans doute une dynamique de l'entropie en sous-jacent qui est à l'œuvre. Cependant, nous allons nous intéresser de nouveau à la Physique Statistique avec un regard "canonique", pour nous faire une idée. Nous aboutirons à des lois plus riches que la loi uniforme.

7.5 Ensemble macro-canonique

Cette notion a été introduite par J. W. Gibbs dans son traité de 1901 (note 5). Il considère un système $\mathscr{S}$ en contact avec un grand réservoir $\mathscr{R}$, sans échange de matière (hypo. de très faible interaction). Typiquement le réservoir a pour fonction de fixer la température. Il y a donc des échanges d'énergie entre $\mathscr{S}$ et $\mathscr{R}$. De plus, on considère que $\mathscr{T} = \mathscr{S} + \mathscr{R}$ est totalement isolé et à l'équilibre. Donc [78]

$$U_{\mathscr{T}} = Cte = U_{\mathscr{R}} + U \tag{213}$$

Si on s'intéresse aux échanges d'énergie δU entre $\mathscr{S}$ et $\mathscr{R}$, il va y avoir des fluctuations qui dépendent de la température [79] T. A l'équilibre, l'entropie totale est maximale et se décompose selon (voir Déf. 1)

$$\mathbb{H}[\mathscr{T}] = \mathbb{H}[\mathscr{R}] + \mathbb{H}, \qquad d\mathbb{H}[\mathscr{T}] = 0 \quad (\textit{équilibre}) \tag{214}$$

et toute chose égale par ailleurs comme volume V, nombre de corpuscules (d), (voir Sec. 4.1.2), nous avons [80]

$$d\mathbb{H}[\mathscr{T}] = \left(\frac{\partial \mathbb{H}[\mathscr{R}]}{\partial U_{\mathscr{R}}}\right)_{x_{\mathscr{R}}}(U_{\mathscr{R}})\ dU_{\mathscr{R}} + \left(\frac{\partial \mathbb{H}}{\partial U}\right)_{x}(U)\ dU \tag{215}$$

Comme $dU_{\mathscr{R}} = -dU$, il vient

$$\left(\frac{\partial \mathbb{H}[\mathscr{R}]}{\partial U_{\mathscr{R}}}\right)_{x_{\mathscr{R}}}(U_{\mathscr{R}}) = \left(\frac{\partial \mathbb{H}}{\partial U}\right)_{x}(U) \tag{216}$$

Cependant, il y a une différence par rapport à la situation de l'équilibre décrite à la section 4.1.2: on a un très grand réservoir, et relativement un petit système. En particulier

78. On laisse tomber l'indice $\mathscr{R}$ pour la notation des grandeurs liées au "petit" système.

79. Rappel: dans le cas micro-canonique T est reliée à la dérivée de l'entropie par rapport à l'énergie du système.

80. NDJE. je note $\left(\frac{\partial f}{\partial x}\right)_{z}(x_o)$ la dérivée partielle de $f(x,z)$ par rapport à x fixant z et évaluée en (x_o, z).

$dU_{\mathscr{T}} \approx dU_{\mathscr{R}} \gg U$. Ainsi,

$$\left(\frac{\partial \mathbb{H}[\mathscr{R}]}{\partial U_{\mathscr{R}}}\right)_{x_{\mathscr{R}}}(U_{\mathscr{R}}) \approx \left(\frac{\partial \mathbb{H}[\mathscr{T}]}{\partial U_{\mathscr{T}}}\right)_{x_{\mathscr{T}}}(U_{\mathscr{T}}) = \frac{1}{T} \tag{217}$$

qui définit la ***température macro-canonique***. En quoi cela conditionne-t-il les états du (petit) système?

On sait que le système total est à l'équilibre micro-canonique, donc tous les états possibles accessibles sont équiprobables. Parmi tous ces états, on compte ceux qui fixent l'énergie du petit système à la valeur U_m, et donc fixent l'énergie du réservoir à la valeur $U_{\mathscr{R}} = U_{\mathscr{T}} - U_m$. Soit $\Omega_{\mathscr{R}}(U_{\mathscr{T}} - U_m)$ l'ensemble des états du réservoir ($|\Omega|$ étant le cardinal), $\Omega(U_m)$ l'ensemble des états du petit système, et $\Omega_{\mathscr{T}}(U_{\mathscr{T}})$ l'ensemble des états pour le tout. La probabilité que le petit système ait une énergie (micro-canonique) U_m est donnée par

$$\mathbb{P}(U_m) = \frac{|\Omega(U_m)||\Omega_{\mathscr{R}}(U_{\mathscr{T}} - U_m)|}{|\Omega_{\mathscr{T}}(U_{\mathscr{T}})|} \tag{218}$$

Or, l'entropie micro $\mathbb{H}$ d'un système est proportionnelle à $\log|\Omega|$, et nous savons que c'est une quantité extensive proportionnelle au nombre de corpuscules présents dans le système. L'hypothèse du réservoir, rend alors son poids statistique bien supérieur à celui du petit système. Ainsi, $\log|\Omega(U_m)|$ est négligeable, et $\log|\Omega_{\mathscr{R}}| \approx \log|\Omega_{\mathscr{T}}|$, et donc [81]

$$\begin{aligned} \log \mathbb{P}(U_m) &\approx \mathbb{H}_{\mathscr{T}}(U_{\mathscr{T}} - U_m) - \mathbb{H}_{\mathscr{T}}(U_{\mathscr{T}}) + Cte \\ &= -U_m \left(\frac{\partial \mathbb{H}_{\mathscr{T}}}{\partial U_{\mathscr{T}}}\right)(U_{\mathscr{T}}) + Cte + \dots \end{aligned} \tag{219}$$

Le premier terme fait apparaitre l'inverse de la température canonique $\beta = 1/T$, donc [82]

$$\mathbb{P}(U_m) \approx Z^{-1}\ e^{-\beta U_m} \tag{220}$$

On trouve ***la distribution de Maxwell-Boltzmann*** dans le cas des vitesses des corpuscules dans un gaz parfait, et Z la constante de normalisation est ce qu'appelle Gibbs ***la fonction***

81. NDJE. les ... sont en premier lieu un terme en U_m/CT avec C la capacité thermique du réservoir qui est très grande par nature.

82. NDJE. il faut entendre T comme $k_B T$ si on veut se replacer dans le contexte de la thermodynamique.

de partition:

$$Z = \sum_{\substack{états \\ U=U_m}} e^{-\beta U_m} \tag{221}$$

Donc, les fluctuations d'énergie (donc des échanges avec le réservoir) du petit système dépendent de la température, c'est-à-dire de la variation d'entropie du réservoir par rapport à l'énergie laquelle ne fluctue pas.

C'est le point de vue de la Physique Statistique, voyons le point de vue mathématique pour comprendre d'où viennent ces lois exponentielles.

7.6 Principe d'entropie maximale

Il s'agit d'un travail entrepris par Edwin Thompson Jaynes (1922-98) en 1957 où il introduit un principe pour tenter de refonder la Mécanique Statistique surtout dans le but d'attaquer des problèmes hors d'équilibre. Jaynes se place dans le contexte de la Théorie de l'Information de Cl. Shannon, où l'entropie est une notion détachée de tout système physique. Le problème posé est de savoir quand on a des observations $\phi_k(x)$ où x représente par une image, un son, etc et ϕ_k peut être n'importe quelle fonction que l'on désire calculée à partir de x, qu'en est-il de la moyenne/espérance?

$$M_n(\phi_k(x)) = \frac{1}{n}\sum_{i=1}^{n} \phi_k(x_i) \tag{222}$$

On sait que si les $(x_i)_i$ sont des *v.a iid*, alors la loi des grands nombres nous dit que M_n converge vers l'espérance. Cependant, que vaut $p(x)$, c'est-à-dire la (densité de) probabilité de x? On aimerait construire un modèle.

Le point de vue de Laplace [83] serait de dire à grand trait: quand on ne sait rien, on peut considérer les observations comme d'égale probabilité. Cela peut s'entendre dans un cas discret fini mais quid d'un cas infini continu. La loi uniforme sur $\mathbb{R}$ pose problème (distribution) et la moyenne empirique ne converge pas. Donc, comment généraliser cette idée "intuitive"? En fait la clé est de remplacer *l'uniformité des probabilités* par ***la maximisation de l'incertitude***. Autrement dit, on va tenter de trouver la probabilité $p(x)$ qui

83. NDJE. « *La probabilité de l'existence d'un événement n'est ainsi que le rapport du nombre des cas favorables à celui de tous les cas possibles, lorsque nous ne voyons d'ailleurs aucune raison pour laquelle l'un de ces cas arriverait plutôt que l'autre.*» extrait de Gérard Jorland https://doi.org/10.4000/ccrh.2772.

diffuse au maximum, ou dit autrement qui impose le moins de contraintes sous-jacentes en dehors des observations. D'où la manifestation de l'entropie. En fait, si l'on impose que pour de 2 v.a indépendantes

$$\mathbb{H}(X,Y) = \mathbb{H}(X) + \mathbb{H}(Y) \tag{223}$$

et que $\mathbb{H}(X) \geq 0$ (cas de valeurs dans un alphabet) Cl. Shannon dans son article de 1948 démontre [84] que l'entropie est égale à une constante près à

$$\mathbb{H} = -\sum_i p_i \log p_i \tag{224}$$

On peut formuler le problème selon: quelle est la probabilité $p(x)$ telle que

1. nous avons une série de moyennes empiriques observées

$$M_n(\phi_k(x)) = \frac{1}{n}\sum_{i=1}^{n} \phi_k(x_i) \tag{225}$$

2. et ***l'entropie*** définie par

$$\mathbb{H}[p] = \begin{cases} -\sum_x p(x)\log p(x) & (discret) \\ -\int p(x)\log p(x)\ dx & (continu) \end{cases} \tag{226}$$

est ***maximum***.

Or, si $x \sim p$ l'espérance de $\phi_k(x)$ est simplement

$$\mathbb{E}_{x\sim p}[\phi_k(x)] = \int \phi_k(x)p(x)dx \tag{227}$$

Donc, nous allons plutôt considérer le problème suivant:

> **Théorème 17** *(**Gibbs**)*

84. Voir le commentaire Cours 2022 Sec. 6.3

S'il existe une distribution de probabilité p telle que

$$(\mu_k = \mathbb{E}_{x\sim p}[\phi_k(x)])_{k\leq K} \text{ (contraintes)}, \quad p = \underset{p}{\text{argmax}}\ \mathbb{H}[p] \tag{228}$$

alors, il existe K paramètres $\Theta = (\theta_k)_{k\leq K}$ tq.

$$p_\Theta(x) = Z_\Theta^{-1} \exp\left\{\sum_k \theta_k \phi_k(x)\right\} = Z_\Theta^{-1}\ e^{\Theta^T \Phi(x)} \tag{229}$$

avec $\Phi(x) = (\phi_k(x))_{k\leq K}$. La constante Z_Θ, ***la fonction de partition de Gibbs****, assure la normalisation de la probabilité, elle prend l'expression*

$$Z_\Theta = \int e^{\Theta^T \Phi(x)} dx \tag{230}$$

Notons qu'alors

$$\mu_k = \mathbb{E}_{x\sim p_\Theta}[\phi_k(x)])_{k\leq K} = \frac{\partial \log Z(\Theta)}{\partial \theta_k} \tag{231}$$

faisant de Z, une pièce maitresse pour les calculs des grandeurs thermodynamiques.

L'on trouve une distribution exponentielle à la Maxwell-Boltzmann. Notons que dans ce contexte

$$U(x) = \Theta^T \Phi(x) \tag{232}$$

et la "température" apparait comme un multiplicateur de Lagrange (au signe près) (ex. θ_0) que l'on peut mettre en facteur, d'où une formulation de l'argument de l'exponentielle $U(x)/T$ (avec le signe approprié selon la convention).

Démonstration 17.

On a à réaliser la maximisation d'une fonction concave $(-p \log p)$, ou la minimisation une fonction convexe, sous contraintes linéaires. S'il existe une solution, elle est unique, que l'on obtient avec les multiplicateurs de Lagrange [85]. On étudie alors la fonction (cas discret)

$$\mathcal{L}(\Theta, p) = -\sum_x p(x) \log p(x) + \sum_{k=1}^K \theta_k \left(\sum_x p(x)\phi_k(x) - \mu_k\right) + \theta_0 \left(\sum_x p(x) - 1\right) \tag{233}$$

85. Cours 2018 Sec. 8.3

La/les variable(s) sont les valeurs des $p(x_i)$ et l'on étudie alors la dérivée (dans le cas continu, il s'agirait d'une dérivée fonctionnelle)

$$\forall x_i, \quad \frac{\partial \mathcal{L}}{\partial p(x_i)} = -1 - \log(p(x_i)) + \sum_k \theta_k \phi_k(x_i) + \theta_0 = 0 \tag{234}$$

d'où $\forall x$ la solution que l'on note p_Θ s'écrit

$$p_\Theta(x) = Z_\Theta^{-1} \exp\left\{ \sum_k \theta_k \phi_k(x) \right\} \tag{235}$$

Mais qu'en est-il de l'adéquation entre p_Θ est la vraie distribution p? Nous avons en fait que [86]

$$D_{KL}(p \| p_\Theta) = \mathbb{H}[p_\Theta] - \mathbb{H}[p] \geq 0 \tag{236}$$

En effet

$$\begin{aligned} D_{KL}(p \| p_\Theta) &= \sum_x p(x) \log \frac{p(x)}{p_\Theta(x)} = -\mathbb{H}[p] - \sum_x p(x) \left(-\log Z + \sum_k \theta_k \phi_k(x) \right) \\ &= -\mathbb{H}[p] + \log Z - \sum_k \theta_k \underbrace{\mathbb{E}_{x \sim p} \left[\phi_k(x) \right]}_{\mu_k} \end{aligned} \tag{237}$$

mais on a également imposé que $\mu_k = \mathbb{E}_{x \sim p_\Theta} \left[\phi_k(x) \right]$ donc on a

$$D_{KL}(p \| p_\Theta) = -\mathbb{H}[p] + \mathbb{E}_{x \sim p_\Theta}[-\log p_\Theta(x)] = \mathbb{H}[p_\Theta] - \mathbb{H}[p] \tag{238}$$

■

Donc finalement, ayant des données, on construit des modèles (paramétrés) de probabilités selon la philosophie de Fisher (Cours de 2022). Cependant, p_Θ n'est pas égale à $p(x)$ que l'on ne connait pas d'ailleurs. Donc en pratique, on définit le meilleur modèle en se servant d'un principe directeur 1) qui satisfait les contraintes observationnelles (les (μ_k)) et 2) le ***principe d'entropie maximale***. On a une liberté: quelles sont les mesures (μ_k) que l'on prend/calcule pour contraindre le modèle? Ex. si on se donne des moyennes et moments du second ordre, on a typiquement un modèle gaussien, mais on peut se

86. NDJE: Attention à l'ordre des arguments de D_{KL}.

donner des fonctions plus complexes, etc, que choisir? On va choisir ce qui réduit le plus l'incertitude, c'est-à-dire $\mathbb{H}[p_\Theta]$ pour s'approcher au mieux de $\mathbb{H}[p]$.

Donc, dans le choix de modèle optimal, on veut trouver les mesures $(\phi_k(x))_k$ ***tq. qu'elles minimisent l'entropie maximum*** $\mathbb{H}[p_\Theta]$**.** C'est un problème de type minimax. Dans le cas de traitement d'images (voir séminaire de Valentin de Portoli 2023), les images de textures peuvent être vues comme un processus aléatoire, dont il faut établir un modèle stochastique. Quelles sont les mesures qu'il faut considérer afin d'établir de bons modèles? Il y a 20-25 ans nous dit Stéphane Mallat, on a essayé des modèles gaussiens de type $U(x) = x^T K x$ où donc $\phi_{k=(i,j)}(x) = x_i x_j$. C'est certes simple mais pas du tout suffisant dans le cas des textures avec de la structure. On peut étendre au cas d'usage de *feature mapping* (Sec. 7.3 Cours 2018), mais surtout ce qui a changé la donne se sont les réseaux de neurones qui fournissent les $\Phi(x)$, c'est-à-dire les représentations optimales. Mais que sont ces $\Phi(x)$, n'a-t-on pas envie de comprendre ce qui est appris par les CNNs? n'a-t-on pas des informations *a priori* qu'il serait bon de mettre dans ces *features* (thème du Cours 2020 entre autres)? Un des enjeux est de ***comprendre les interactions entre les différentes échelles du problème***.

8. Séance du 1er Mars

Nous allons revenir sur la conception de modèles guidée par ***le principe d'entropie maximale*** et par les ***mesures disponibles*** $((\mu_k)_{k\leq K})$ obtenues comme espérances de fonctions des données $((\mathbb{E}[\phi_k(x)])_k)$. Le théorème de Gibbs (Th. 17) nous donne accès à une approximation de $p(x)$, la densité de probabilité, selon un ***modèle paramétrique exponentiel*** $p_\theta(x)$ où ***les paramètres*** Θ ***sont les multiplicateurs de Lagrange associés aux contraintes des mesures***, et sont donc les ***variables duales des mesures***. Nous allons étudier cette dualité, car rappelons-nous, ces paramètres qui pourraient paraitre des abstractions mathématiques, sont des grandeurs bien concrètes en Physique comme la température, la pression, la viscosité, etc.

Dans le processus de maximisation de l'entropie, nous trouvons $p_\Theta(x)$ qui approxime $p(x)$ de telle façon que la divergence entre les deux distributions est donnée par

$$D_{KL}(p\|p_\Theta) = \mathbb{H}(p_\Theta) - \mathbb{H}(p) \geq 0 \tag{239}$$

Ainsi, l'erreur est due à l'excès d'entropie, c'est-à-dire un trop d'incertitude qui nous a manqué pour construire le modèle. Donc, le problème auquel on est confronté concerne ***le choix des mesures*** $\phi_k(x)$ ***qui contraignent le modèle*** (c'est-à-dire les paramètres Θ). On aimerait donc minimiser $D_{KL}(p\|p_\Theta)$ tout en maximisant l'entropie. Et cela se fait en trouvant ***les mesures les plus informatives***.

8.1 Exemple: la distribution gaussienne

Quelles sont donc les mesures que l'on peut effectuer? Quand on a une *v.a*[87] $X = (X_1, \dots, X_d)^T \in \mathbb{R}^d$, on peut vouloir estimer la moyenne, et également les moments d'ordre 2

$$\mu = \mathbb{E}_p(X) = (\mathbb{E}_p(X_i))_{i\leq d} \qquad \Sigma = \mathbb{E}_p(XX^T) = (\mathbb{E}_p(X_iX_j))_{(i,j\leq d)} \tag{240}$$

Quelle est l'expression de $p_\Theta(x)$? D'après le théorème 17, alors

$$p_\Theta(x_1, \dots, x_d) = Z_\Theta^{-1} \exp\left\{\sum_{i=1}^d \theta_i x_i + \sum_{i,j=1}^d \theta_{ij} x_i x_j\right\} \tag{241}$$

87. NDJE. Pour commencer juste un peu plus simplement, prenons une variable $X \in \mathbb{R}$ dont on connait, l'espérance μ et la variance σ^2. Les contraintes se traduisent selon $\mathbb{E}[X] = \mu$ et $\mathbb{E}[X^2] = \sigma^2 + \mu^2$, d'où $\phi_1(x) = x$ et $\phi_2(x) = x^2$. Alors on trouve que

$$p_\Theta(x) = Z^{-1} e^{\theta_1 x + \theta_2 x^2} \qquad Z_\Theta = \int e^{\theta_1 x + \theta_2 x^2} dx$$

Or,

$$\mu = \frac{\partial \log Z}{\partial \theta_1} \quad \text{et} \quad \sigma^2 + \mu^2 = \frac{\partial \log Z}{\partial \theta_2}$$

De plus, Z a pour expression ($\theta_2 < 0$)

$$Z(\theta_1, \theta_2) = \left(\frac{\pi}{\theta_2}\right)^{1/2} e^{\theta_1^2/(4\theta_2)}$$

Ainsi, on trouve que $1/\theta_2 = -2\sigma^2$, $\theta_1 = \mu/\sigma^2$ et $Z = \sqrt{-\pi/\theta_2} e^{-\theta_1^2/(4\theta_2)}$. Finalement, la loi $p_\Theta(x)$ prend la forme

$$p_\Theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} = \mathcal{N}(\mu, \sigma^2)$$

Le cas traité par S. Mallat est une généralisation où $X \in \mathbb{R}^d$.

On peut "compléter les carrés" et en exprimant les θ en fonction de μ on obtient l'expression

$$p_\Theta(x) = Z_\Theta^{-1} \exp\Big\{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\Big\} \tag{242}$$

où la matrice C est la covariance (l'espérance du moment d'ordre 2 centré)

$$C = \mathbb{E}_p((X-\mu)(X-\mu)^T) = \Sigma - \mu\mu^T \tag{243}$$

et $Z_\Theta = ((2\pi)^d |\det C|)^{1/2}$. On vérifie en effet que μ est bien la moyenne de p. Concernant C, étant une matrice symétrique donc diagonalisable, l'on peut opérer des rotations afin de se placer dans la base où C est diagonale, ce qui permet de calculer les moments d'ordre 2 par produits d'intégrales.

Ce que cela nous dit, c'est que ***la distribution gaussienne peut être interprétée comme la distribution d'entropie maximum contrainte par des moyennes et des moments du second ordre***. Maintenant, les distributions gaussiennes sont certes très pratiques et peuvent suffirent dans certains cas, mais ***elles ne sont pas capables de capturer les structures*** d'un champ turbulent ou des textures. Donc quelles sont les ϕ_k qu'il nous faut ajouter au delà des moments d'ordre 2?

Dans le théorème 17, il y a clairement ***un lien de dualité entre les observables et les multiplicateurs de Lagrange***

$$(\mu_k = \mathbb{E}_{x\sim p}[\phi_k(x)])_{k\leq K} \longleftrightarrow (\theta_k)_{k\leq K} \tag{244}$$

La dualité nous dit qu'en fait la représentation de p_Θ peut s'écrire avec l'un ou l'autre des jeux de variables (voir par ex. la note 87). Nous allons approfondir ce lien car cette notion de dualité est fondamentale, non seulement en Mathématiques dans les problèmes d'optimisation, mais aussi en Physique à travers les équations de Lagrange et d'Hamilton. Tout ceci est relié par la transformation de Legendre. Avant cela voyons les propriétés de base de la fonction de partition.

8.2 Fonction de partition Z_Θ

Nous avons vu dans le théorème 17 que les observables μ_k et le multiplicateur de Lagrange associé θ_k sont reliés via la fonction de partition selon

$$\mu_k = \frac{\partial \log Z(\Theta)}{\partial \theta_k} \tag{245}$$

Définissons alors la fonction, nommée *énergie libre* en Physique $F(\Theta)$ et en Optimisation $A(\Theta)$ *la fonction cumulante* selon

$$A(\Theta) = -F(\Theta) = \log Z(\Theta) \tag{246}$$

Théorème 18 ***(fonction cumulante)***
Dans le cadre du théorème 17, si la fonction $A(\Theta) = \log Z_\Theta$ a des dérivées d'ordres supérieurs sur son support Λ, alors primo

$$\nabla_\theta A(\Theta) = \mu = \mathbb{E}(\Phi(x)) \qquad \text{cad. } \forall k \leq K,\ \mu_k = \frac{\partial A(\Theta)}{\partial \theta_k} = \mathbb{E}(\phi_k(x)) \tag{247}$$

et secundo si l'on regarde le Hessien alors

$$\nabla^2_\theta A(\Theta) = Cov(\Phi(x)) \geq 0 \qquad \text{cad. } \forall (k,k') \leq K, \frac{\partial^2 A(\Theta)}{\partial \theta_k \partial \theta_{k'}} = Cov(\phi_k(x)\phi_{k'}(x)) \geq 0 \tag{248}$$

(nb. les espérances sont prises par rapport à la distribution $p_\Theta(x)$)

La démonstration ne pose pas de problème, il suffit de dériver sous le signe intégrale et donc il faut être dans les conditions pour le faire, mais avec une exponentielle ça se passe bien. Ce théorème nous dit que c'est bien la fonction de partition qui va caractériser tous les moments d'ordres supérieurs de p_Θ.

Faisons alors quelques observations:

- Le Hessien de A est positif (c'est une matrice de covariance) donc, A ***est une fonction convexe***.

- Ensuite, soit la définition suivante:

> **Définition 18** *(**famille libre**)*
> *On qualifie* $(\phi)_{k\leq K}$ *de* ***minimum****, si c'est une* ***famille linéairement indépendante****:*
> $$\sum_k \beta_k \phi_k = B^T\Phi = \Phi^T B = 0 \Leftrightarrow B = 0 \tag{249}$$

Dans ce cas $C = Cov(\Phi(x)) > 0$ c'est-à-dire qu'***il n'y a aucun vecteur propre qui soit associé à la valeur propre nulle***. En effet

$$\begin{aligned}\forall B \in (\mathbb{R}^K \backslash \{0\}),\ \mathbb{E}[(B^T(\Phi(x)-\mu))^2] &= \mathbb{E}[B^T(\Phi(x)-\mu)(\Phi(x)-\mu)^T B] \\ &= B^T\mathbb{E}[(\Phi(x)-\mu)(\Phi(x)-\mu)^T]B \\ &= B^T C B = f(B) > 0 \end{aligned} \tag{250}$$

Or, $f(B)$ étant une forme quadratique strictement positive, alors les valeurs propres de C sont *strictement* positives. Cela se montre en faisant remarquer que C étant symétrique réelle, alors il existe une matrice orthogonale Q et une matrice diagonale D telle que $Q^T C Q = D = diag((\lambda_k)_{k\leq K})$, si $\tilde{B} = QB$ alors

$$B^T C B > 0, \forall B \neq 0 \Leftrightarrow \tilde{B}^T D \tilde{B} > 0, \forall \tilde{B} \neq 0 \Leftrightarrow \sum_{k=1}^{K} \lambda_k \tilde{b}_k^2 > 0, \forall \tilde{B} \neq 0 \tag{251}$$

Si on prend comme $\tilde{B}$ les vecteurs de la base canonique de $\mathbb{R}^K$ alors on conclut que toutes les $v.p$ de C sont strictement positives.

Dans ce contexte où $(\phi(x)_k)$ est une ***famille libre***, alors *le Hessien* qui est égal à la matrice de covariance, est *strictement positif*. ***Le problème est alors strictement convexe*** ce qui est particulièrement intéressant dans l'usage de la descente de gradient. En particulier,

$$\nabla^2 A(\Theta) > 0 \Leftrightarrow (\nabla A(\Theta_1) - \nabla A(\Theta_2)).(\Theta_1 - \Theta_2) > 0 \tag{252}$$

et donc il ne peut y avoir 2 jeux de paramètres Θ qui redonnent les contraintes μ_k. ***Il y a unicité de la solution***, c'est-à-dire qu'il n'y a qu'un unique jeu de multiplicateurs de Lagrange.

8.3 Dualité conjuguée: transformée de Legendre-Fenchel

Remarquons qui si l'on dispose des valeurs des multiplicateurs de Lagrange optimaux Θ^*, nous disposons alors de la densité de probabilité p_{Θ^*}, et non seulement les $(\mu_k)_k$ sont accessibles, mais aussi tout autre moment d'ordre supérieur de la distribution. Ceci étant dit, comment calculer le vecteur Θ à partir des valeurs des $(\mu_k)_k$? C'est alors que la transformée de Legendre va apparaitre et nous allons porter un autre regard sur ce que nous sommes en train de faire qui revêt un caractère plus général.

Plaçons nous dans un cadre d'optimisation convexe général. On peut reparamétriser la fonction, et travailler dans espace dual. Introduisons pour ce faire la transformation de Legendre-Fenchel [88].

Définition 19 *(Legendre-Fenchel)*
Soit une fonction convexe $f : \chi \to \mathbb{R}$ ou $\bar{\mathbb{R}}$ (contient l'infini), on définit sa fonction conjuguée par la transformation de Legendre-Fenchel selon [a]

$$L[f](s) = \sup_{x \in \chi} \{s^T x - f(x)\} \tag{253}$$

La solution du problème est que $s = \nabla_x f(x)$ ce qui fournit implicitement une relation entre x et s.

a. nb. il y a plusieurs conventions concernant le signe global de la transformation, et sur le signe relatif entre les deux éléments.

(NDJE. Si la transformée de Legendre-Fenchel est d'abord définie pour des fonctions convexes, et c'est en cela quelle est utilisée ci-après, son application pour une fonction non convexe est importante afin d'obtenir une enveloppe convexe de celle-ci. C'est une propriété utilisée dans la théorie de Landau des transitions de phases. Voir la section 8.8 pour un exemple de convexification d'un chapeau mexicain.)

88. NDJE. La transformation porte les noms d'Adrien-Marie Legendre (1752-1833) qui l'utilisait en Mécanique Analytique pour le passage du Lagrangien à l'Hamiltonien et en Thermodynamique; et de Moritz Werner Fenchel (1905-88) connus en particulier pour ses travaux en analyses convexes. Je prends la notation $L[f]$ car $*$ est souvent utilisé pour désigné le résultat d'une optimisation comme un argmax. Mais dans la littérature vous trouverez f^*.

Propriété 6 *(**convexité de** $L[f]$)*
La transformée de Legendre-Fenchel d'une fonction convexe est une fonction convexe.

Démonstration 6. On va juste montrer [89] cette propriété dans le cas où f est convexe et doublement différentiable (1D) et $f'' > 0$. Une solution au problème s'écrit donc $s = f'(x)$. Notons par hypothèse f' est strictement monotone et inversible. Donc, si on note $g = (f')^{-1}$, $x = g(s)$, et donc $f'(g(s)) = s$. La fonction g est aussi différentiable $g'(s) = 1/f''(g(s))$. Et $L[f](s) = g(s)s - f(g(s))$ est aussi différentiable avec

$$L[f]'(s) = g(s) + g'(s)(s - \underbrace{f'(g(s))}_{s}) = g(s)$$

et donne

$$L[f]''(s) = g'(s) = 1/f''(g(s)) > 0$$

d'où la convexité de la transformation de Legendre dans ce cas mais on peut généraliser la démonstration [90]. ■

La transformation de Legendre $L[f]$ est également appelée ***la conjuguée convexe de*** f. La relation $s = \nabla_x f(x)$ nous permet d'envisager de paramétriser f non pas en fonction de x mais en fonction de s.

Théorème 19 *(**synthèse des fonctions convexes**)*
*Si f est **convexe** alors on peut la synthétiser à partir de la transformation de Legendre appliquée à $L[f]$:*

$$L[L[f]](x) = L^2[f](x) = \sup_s \{s^T x - L[f](s)\} = f(x) \tag{254}$$

Là encore une solution nous est fournie par $x = \nabla_s L[f](s)$. Dans ce cas $L^2 = Id$ (involution).

89. NDJE. C'est une propriété que je rajoute au cours car S. Mallat n'a pas pu en parler faute de temps.
90. NDJE. Voir par exemple https://www.math.univ-toulouse.fr/~weiss/Docs/LectureNotes_ConvexOptimization_PWeiss.pdf.

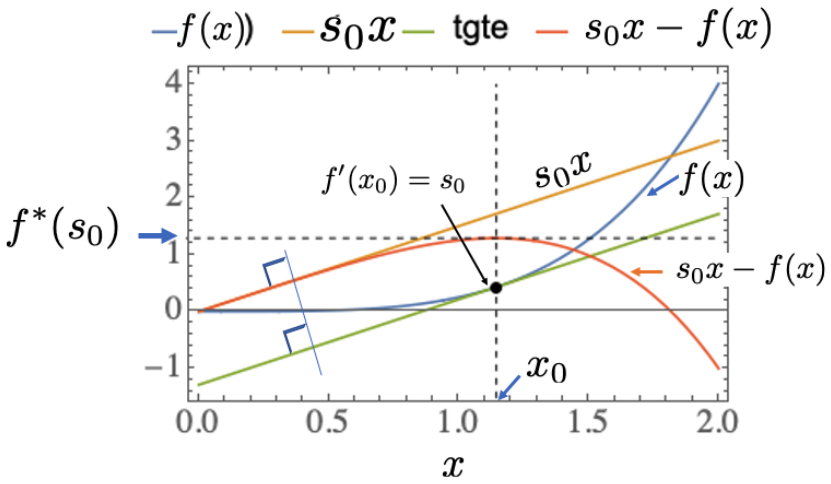


FIGURE 26 – Illustration de la transformation de Legendre-Fenchel de la fonction $f(x)$ pour donner la fonction $L[f](s)$. Pour un s_0 donné, le maximum de $s_0x - f(x)$ est atteint pour x_0 tel que $f'(x_0) = s_0$. On a donc une paramétrisation de x en fonction de s que l'on peut exploiter pour transformer $f(x)$.

NDJE. Un exemple de fonction convexe: $f(x) = \frac{1}{4}x^4$, alors trouver s à partir de x donne $s = f'(x) = x^3$, donc ici le cas est simple, $L[f](s) = \frac{3}{4}s^{4/3}$. Inversement, $x = (L[f])'(s) = s^{1/3}$ donc, on trouve que $L[L[f]](x) = x^{4/3} - 3/4x^4 = f(x)$. On retrouve alors $f(x)$ en appliquant deux fois la transformation de Legendre.

Un exemple en 1D est donné sur la figure 26. Il y a une interprétation géométrique à $L[f](s)$: le maximum de $sx - f(x)$ est atteint pour la valeur de x_0, et le point $(x_0, f(x_0))$ est celui où la tangente du graphe de f en ce point est de pente égale à s_0.

Une application en Mécanique Analytique est constituée par le passage du Lagrangien [91] $\mathscr{L}(q, \dot{q}, t)$ qui donne avec le principe de moindre action les équations d'Euler-Lagrange, à l'hamiltonien $H(q, p, t)$ défini par la transformation de Legendre

$$p(q, \dot{q}, t) := \frac{\partial \mathscr{L}(q, \dot{q}, t)}{\partial q} \qquad H(q, p, t) = \dot{q}\, p - \mathscr{L}(q, \dot{q}, t) \tag{255}$$

91. NDJE, ici une seule particule est considérée mais cela se généralise aisément. La notation $\dot{q}$ indique la dérivée de $q(t)$ par rapport au temps t.

qui donne les équations d'Hamilton où la force newtonnienne apparait comme la dérivée temporelle de p. Une autre application en Physique Statistique est le passage de l'énergie à l'énergie libre.

Maintenant, si on inspecte le théorème 18, on se rend compte que la relation

$$\mu = \nabla_\Theta A(\Theta) \tag{256}$$

joue le même rôle que $s = \nabla f(x)$. On peut donc écrire

$$L[A](\mu) = \sup_\Theta \{\mu^T \Theta - A(\Theta)\} \tag{257}$$

$$L^2[A](\Theta) = \sup_\mu \{\mu^T \Theta - L[A](\mu)\} \tag{258}$$

Voyons le lien entre la transformation de Legendre et l'optimisation des Θ.

8.4 Optimisation de Θ en fonction de μ

Notre but initial est de trouver le modèle paramétrique $p_\Theta(x)$ s'approchant au mieux de $p(x)$ en ayant les contraintes μ (mesures) et suivant le principe d'entropie maximum. Si l'on dispose de n échantillons $(x_i)_{i\leq n}$ supposés tirés aléatoirement de $p(x)$ (iid), alors

$$\mathbb{E}_p[\log p_\Theta] = \int p(x) \log p_\Theta(x)\ dx \approx \frac{1}{n} \sum_{i=1}^{n} \log p_\Theta(x_i) \tag{259}$$

Le principe du ***Maximum de Vraisemblance***[92] de R. Fisher, nous dit alors que si $p_\Theta(x)$ est réellement un bon modèle de $p(x)$ et que x_i est un exemple typique de $p(x)$; alors on s'attend aussi à ce que $p_\Theta(x_i)$ soit également grand, et donc la moyenne aussi.

Dans le cas où p_Θ à la forme d'une distribution de Gibbs alors

$$\log p_\Theta(x) = \Theta^T \Phi(x) - \log Z(\Theta) = \Theta^T \Phi(x) - A(\Theta) \tag{260}$$

Donc pour obtenir le Θ optimal, aussi appelé Maximum Likelihood Estimator (MLE), il

92. Voir Cours 2022 Sec. 3.5

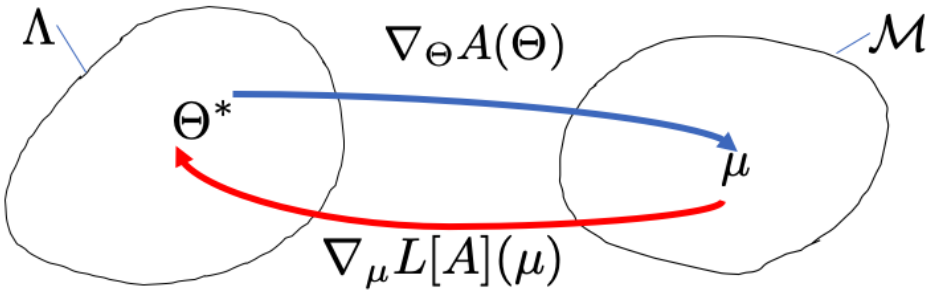


FIGURE 27 – Illustration de la relation duale entre le jeu de mesures μ et le jeu de paramètres Θ.

nous faut réaliser

$$\max_\Theta \mathbb{E}_p[\log p_\Theta] = \max_\Theta\{\Theta^T\mathbb{E}_p[\Phi(x)] - A(\Theta)\} = \max_\Theta\{\Theta^T\mu - A(\Theta)\} = L[A](\mu) \quad (261)$$

On trouve alors le lien avec la transformation de Legendre de la fonction $A(\Theta)$ avec une correspondance $s \leftrightarrow \mu$ et $x \leftrightarrow \Theta$ du cas général de la section précédente (Déf. 19). Ainsi, ayant les mesures μ, le Θ^* optimal satisfait

$$\nabla_\Theta A(\Theta^*) = \mu \quad (262)$$

Mais, on sait que $L[A]$ est convexe (d'ailleurs indépendamment de la convexité de A) et donc

$$\Theta^* = \nabla_\mu L[A](\mu) \quad (263)$$

On peut alors résumer la relation entre μ et Θ si l'on se place dans le cadre d'une famille libre de mesures (Déf. 18) sur le schéma 27. Dans un cas il nous faut calculer $A(\Theta)$ pour obtenir son gradient, et dans l'autre cas c'est $L[A](\mu)$ qui nous occupe, et on doit maximiser la vraisemblance pour obtenir Θ^* optimal.

Pour obtenir Θ^* il nous en définitive réaliser

$$\Theta^* = \operatorname*{argmax}_\Theta\ (\Theta^T\mu - A(\Theta)) = \operatorname*{argmin}_\Theta\ (A(\Theta) - \Theta^T\mu) \quad (264)$$

Typiquement, en optimisation on peut réaliser une descente de gradient [93], donc entre les paramètres obtenus aux étapes t et $t+1$, nous avons la relation suivante:

$$\begin{aligned}\Theta_{t+1} - \Theta_t &= -\nabla_\Theta\{A(\Theta) - \Theta^T\mu\}\Big|_{\Theta=\Theta_t} = -\nabla_\Theta A(\Theta_t) + \mu \\ &= \mu - \mu(\Theta_t) \end{aligned} \tag{265}$$

où d'après la paramétrisation de Gibbs (Th. 17)

$$\mu(\Theta) = \mathbb{E}_{x\sim p_\Theta}[\Phi(x)] \tag{266}$$

Si donc on part d'une valeur Θ_0 quelconque, alors à partir de la distribution p_{Θ_0} on peut calculer les moments $\mu(\Theta_0)$ à l'aide des espérances. Il est clair qu'au début ces moments ne seront pas égaux aux vrais moments μ qui par hypothèse sont donnés par

$$\mu = \mathbb{E}_{x\sim p}[\Phi(x)] \tag{267}$$

Donc, il y a une erreur entre μ et $\mu(\Theta_0)$, et la nouvelle valeur Θ_1 doit donc aller contre le gradient et d'une quantité égale en intensité à l'erreur sur les moments. On itère la procédure, et comme ***nous sommes dans un cas strictement convexe*** (voir les observations faites à la suite du théorème 18) alors ***la solution existe et est unique***.

Donc, ***la relation entre le principe du Maximum d'Entropie et le principe du Maximum de Vraisemblance, est une relation duale que l'on comprend par le biais de la Transformée de Legendre-Fenchel du logarithme de la fonction de partition*** $Z(\Theta)$ qui soit se nomme énergie libre F en Physique Statistique ou fonction cumulante A en Optimisation.

Finalement, le cadre mathématique décrit ci-dessus montre le caractère général, et sort la notion de *fonction de partition* et *d'énergie libre* de leur cadre de la Physique Statistique où Gibbs les a introduit au début du XXe siècle. Mais également dans le cas d'optimisation, on peut voir le problème comme celui de la maximisation de l'entropie sous contraintes (observables, μ), ou celui du maximum de vraisemblance.

93. Voir Cours 2018 Sec. 10.1 par exemple.

8.5 Problème de l'estimation de la qualité de Θ_t

Si l'on regarde un peu plus soigneusement l'équation 265 de la descente de gradient pour passer de Θ_t à Θ_{t+1}, il nous faut pouvoir tester cette nouvelle valeur des paramètres. Pour ce faire, il faut pouvoir calculer les nouvelles estimations des moments. Donc, d'une manière générale, il nous faut pouvoir calculer $\forall k$

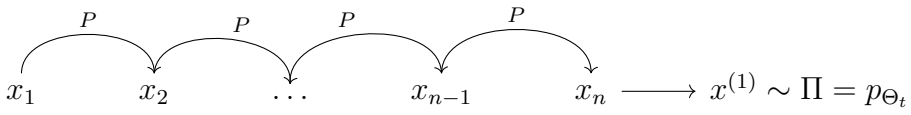
$$\int \phi_k(x)\ p_\Theta(x)\ dx \tag{268}$$

Comment peut-on faire ce type de calcul d'intégrale? Une méthode possible qui paraît naturelle est la suivante

1. tirez N réalisations $x \sim p_\Theta(x)$;
2. calculer la moyenne $1/N \sum_{i=1}^{N} \phi_k(x_i)$. L'erreur est en $1/\sqrt{N}$.

Le hic est à l'étape (1) car ***il nous faut pouvoir échantillonner*** $p_\Theta(x)$ ce qui pose déjà des problèmes en faible dimension si la distribution à des modes multiples, mais cela devient de plus plus en plus difficile en grande dimension même dans un cas où il n'y a qu'un seul mode.

Une technique très utilisée est celle de la production de chaines de Markov (Sec. 6.4) que l'on appelle: ***Monte Carlo Markov Chain*** (MCMC). On construit des ***chaines de Markov dont la mesure invariante (Déf. 11) est précisément*** p_Θ**.** C'est-à-dire pour produire 1 échantillon x_i tiré selon p_{Θ_t}, pour déterminer les moments $\mu_k(\Theta_t)$, il nous faudra faire évoluer une chaine de Markov selon une matrice de transition telle que la loi invariante soit p_{Θ_t}



Puis on réitère le processus pour obtenir $x^{(2)}$ et ainsi de suite pour obtenir suffisamment d'échantillons afin de calculer les moyennes et estimer la qualité de $\mu(\Theta_t)$.

En pratique, il y a plusieurs algorithmes qui produisent de telles chaines comme celui de ***Metropolis-Hastings*** *(NDJE. Voir section 8.9 pour une introduction. Vous pouvez expérimenter avec les notebooks de 2023 https://github.com/jecampagne/cours_mallat_cdf/tree/main/cours2023*). Nous ne rentrons pas plus en détail dans les algorithmes MCMC.

8.6 Comment concevoir des $\Phi(x)$?

Au bilan des sections précédentes, ***si l'on dispose de*** $(\phi_k)_k$ ***bien adaptés au problème, on a un formalisme qui se déroule bien***. Surtout quand on est dans le cas convexe. ***Mais le problème reste cependant posé: que sont les*** $(\phi_k)_k$, si on a des images de visages, de textures, des trames sonores de musiques, de locuteurs, etc, qui vont nous permettre d'inférer efficacement une distribution de probabilité de Gibbs (exponentielle), laquelle permettra de générer de nouvelles images, textures, de nouvelles trames sonores, etc, par échantillonnage. Pour répondre à ce problème il nous faut revenir à la base et se poser la question: ***quelles sont les propriétés que l'on peut identifier et intégrer à la forme des*** $\phi(x)$? Il s'agit de la thématique de l'***information a priori*** visitée par exemple dans le cours de 2020.

Maintenant, comment exprimer cette information *a priori* en termes mathématiques exploitables. Même si l'on peut disposer de descriptions expérimentales précises, par exemple d'une image de fluide turbulent, ou d'une texture d'écorce d'arbre, etc, ***le problème est de concrétiser en termes mathématiques qui définissent les*** $\phi(x)$***: là est le point difficile***.

Ce qui a fonctionné jusqu'à présent dans le déroulement de la Physique (Statistique, Particules élémentaires et Relativité) et en Mathématiques également, c'est de regarder ***les symétries du problème***.

8.7 Symétries de $p(x)$

La question à laquelle on s'attaque est de connaître les transformations de x qui laissent invariante $p(x)$. Typiquement soit g une telle transformation, en général un élément d'un groupe G, qui agit sur x pour le transformer en x' de telle façon que

$$g \in G, x' = g.x \xRightarrow{invariance} p(x') = p(x) \tag{269}$$

G est en effet un groupe car

- $g = Id$ est manifestement l'élément neutre de G;
- si $(g_1, g_2) \in G$,

$$p((g_1 \circ g_2).x) = p(g_1.(g_2.x)) = p(g_2.x) = p(x)$$

donc $g_1 \circ g_2 \in G$;

- on peut finalement imposer que g ait une inverse telle que $g \circ g^{-1} = Id$.

Le groupe G le plus classique est celui des ***translations***. Dans ce cas si l'on conçoit que x dépend d'une variable u, par ex. $u = (u_1, u_2)$ la position d'un pixel dans une image et $x(u)$ l'intensité du signal déposé, alors l'action d'une translation de "vecteur" g s'exprime

$$x'(u) = x(u - g) = g.x(u) \tag{270}$$

et ***la stationnarité*** s'exprime selon

$$p(x) = p(g.x) \Leftrightarrow \forall u,\ p(x(u)) = p(x(u - g)) \tag{271}$$

La stationnarité est la manifestation de la symétrie par rapport au groupe des translations. Voilà un exemple où l'on passe d'un constat expérimental (*"stationnarité"*) à une caractérisation mathématique (*"symétrie par rapport au groupe des translations"*).

Bien entendu, il peut y avoir d'autres types de symétries: ex. ***groupe des rotations***. Mais la symétrie de rotation est peut-être adaptée pour certains cas de figures comme en Cosmologie quand on regarde le Ciel, mais pas par exemple dans le cas d'images de visages où la verticalité impose une direction privilégiée. On peut parfois également avoir des ***invariances par changements d'échelle***: ex. en traitement d'image, la distance entre l'objet et la caméra ne change pas la nature de l'objet, et donc la collection de toutes les images possibles de l'objet en question a une distribution de probabilité invariante d'échelle. Mais par contre, les photos d'identités ne sont pas invariantes par dilatation car le pré-traitement fait en sorte de recadrer la prise de vue, afin de fournir une image acceptable pour un passeport ou une CNI. Tout cela pour dire que ***chaque problème a ses symétries propres***. Ce qui marche dans un cas/une application peut très bien ne pas fonctionner pour un/une autre. Maintenant, on peut envisager des symétries complexes (ex. les déformations locales ou difféomorphismes).

Mais en quoi cela sert-il de connaitre les symétries de $p(x)$? La principale raison vient de ***la réduction de la dimensionalité*** du problème quelles nous permettent d'opérer. Cela se fait par le biais de la réduction du nombre K de mesures $(\phi_k(x))_{k \leq K}$, en imposant des invariants, et donc de ce fait cela implique également une réduction du nombre de paramètres $(\theta_k)_{k \leq K}$ du modèle $p_\Theta(x)$ s'approchant au mieux de $p(x)$.

Notons que l'on a K mesures et que x est de dimension d ($x \in \mathbb{R}^d$): la question que l'on pourrait se poser est quelle est la relation entre K et d? Par un raisonnement développé à la section 2.6, on s'attend à ce que K évolue comme l'exponentielle de d: il s'agit de la ***malédiction de la dimensionalité***. Donc, K devrait être colossal dès lors que $d \sim 10^6$. Mais si l'on veut pouvoir estimer $p(x)$ à partir d'une seule image, alors $K \ll d$. ***Cela signifie qu'il faut énormément d'information a priori pour opérer une telle réduction de dimensionalité***. C'est là où les symétries vont aider.

Définition 20 *(**invariant linéaire**)*
Soit un groupe fini G de transformations linéaires[a] *de x et soit **l'opérateur de moyenne**, noté $\mathcal{M}$ défini selon*

$$\mathcal{M}.x = \frac{1}{|G|} \sum_{g \in G} g.x \tag{272}$$

C'est-à-dire que l'on opère une moyenne sur les éléments de l'orbite de x, soit l'ensemble $O_x = \{g.x; g \in G\}$. (nb. on peut généraliser à un groupe continu).

a. ex. translation, rotation, dilatation

$\mathcal{M}$ est un opérateur invariant par l'action du groupe. En effet: pour tout $g_0 \in G$

$$\mathcal{M}.(g_0.x) = \frac{1}{|G|} \sum_{g \in G} g.(g_0.x) = \frac{1}{|G|} \sum_{g \in G} (g \circ g_0).x \underbrace{=}_{g' = g \circ g_0^{-1}} \frac{1}{|G|} \sum_{g' \in G} g'.x = \mathcal{M}.x \tag{273}$$

Dans le cas d'une image et G le groupe des translations de paramètres τ

$$\mathcal{M}.x = \sum_{\tau} x(u - \tau) \tag{274}$$

qui n'est rien d'autre que la moyenne du signal contenu dans l'image, lequel est bien un invariant [94].

La conséquence est que l'on peut se restreindre à des $\phi(x)$ qui sont invariants.

94. NDJE: pour s'en convaincre prenons un signal $x(t)$ périodique sur $[0, 2\pi]$, simplement la somme sur θ se traduit par une intégrale $(2\pi)^{-1} \int_0^{2\pi} x(u - \theta) d\theta$ qui par changement de variable $u' = u - \theta$ et périodicité du signal, donne $(2\pi)^{-1} \int_0^{2\pi} x(t) dt$ qui n'est autre que la moyenne du signal.

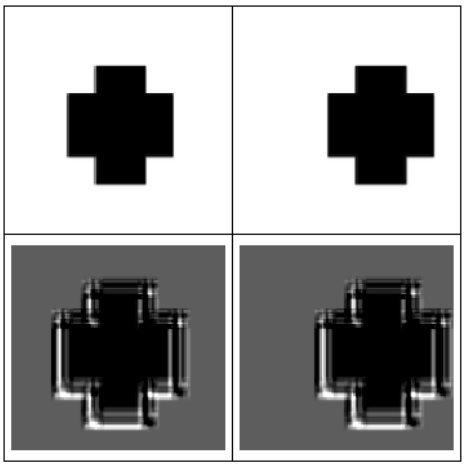


FIGURE 28 – Illustration de la propriété d'équivariance de la convolution par application d'une translation: (haut) de gauche à droite: image origine, image translatée; (bas) de gauche à droite: convolution de l'image d'origine par un filtre, et convolution par le même filtre de l'image translatée. La convolution de l'image translatée, et la translatée de la convolution de l'image d'origine. Si x est l'image, f la convolution et g la translation alors $f(g.x) = g.f(x)$.

Définition 21 *(**équivariance**)*
On dit que $\phi(x)$ est équivariante par l'action du groupe G, si

$$\forall g \in G,\ \phi(g.x) = g.\phi(x) \tag{275}$$

c'est-à-dire que ϕ commute avec les éléments de G.

Une illustration de la propriété d'équivariance de la convolution par action du groupe des translations est proposée sur la figure 28.

Notons que l'*équivariance* ou autrement appelée *covariance* se distingue de l'*invariance*:

$$\begin{cases} f(g.x) = f(x) & \text{invariance} \\ f(g.x) = g.f(x) & \text{équivariant/covariance} \end{cases} \tag{276}$$

Nous avons un résultat immédiat:

Propriété 7 *Dans le cas où donc ϕ est équivariante, l'opérateur*

$$\tilde{\phi}(x) = \mathcal{M}.\phi(x) = \frac{1}{|G|}\sum_{g\in G} g.\phi(x) = \frac{1}{|G|}\sum_{g\in G}\phi(g.x) \tag{277}$$

est ***invariant***.

En effet, par le même type de changement de variable que précédemment,

$$\tilde{\phi}(g'.x) = \frac{1}{|G|}\sum_{g\in G}\phi(g.(g'.x)) = \frac{1}{|G|}\sum_{g\in G}\phi(g.x) = \tilde{\phi}(x) \tag{278}$$

Théorème 20

Si Φ est équivariant sur G, et si $p(x)$ est invariant selon G alors nous avons les propriétés suivantes

1) $\mu = \mathbb{E}_p(\Phi(x)) = \mathbb{E}_p(\mathcal{M}.\Phi(x))$
2) *la distribution d'entropie maximum $p(x)$ contrainte par les mesures $\mathbb{E}_p(\Phi(x)) = \mu$ s'écrit*

$$p(x) = Z^{-1}e^{-\Theta^T\mathcal{M}\Phi(x)} \tag{279}$$

Démonstration 20. Pour la première propriété, on peut la démontrer en développant l'expression $\mathbb{E}_p(\mathcal{M}.\Phi(x))$ pour un des éléments de $\Phi = (\phi_k)_{k\leq K}$ que l'on écrit ϕ pour

alléger. Il vient:

$$\begin{aligned}
\mathbb{E}_p(\mathcal{M}.\phi(x)) &= \sum_{x\in\chi} p(x)\frac{1}{|G|}\sum_{g\in G}\phi(g.x) && (\textit{équivariance de } \phi)\\
&= \frac{1}{|G|}\sum_{g\in G}\sum_{x\in\chi} p(x)\phi(g.x)\\
&= \frac{1}{|G|}\sum_{g\in G}\sum_{x'\in\chi} p(g^{-1}.x')\phi(x') && (x' = g.x)\\
&= \frac{1}{|G|}\sum_{g\in G}\left(\sum_{x\in\chi} p(x)\phi(x)\right) && (\textit{invariance de } p)\\
&= \sum_{x\in\chi} p(x)\phi(x) && (\sum_{g\in G} a = |G|a)\\
&= \mathbb{E}_p(\phi(x)) && (280)
\end{aligned}$$

La seconde propriété découle du fait que l'on peut chercher $p_\Theta(x)$ selon le théorème de Gibbs (Th. 17) avec les contraintes $\mu = \mathbb{E}_p(\mathcal{M}.\Phi(x))$. ■

Cela nous dit que ***lorsque l'on moyenne, on réduit le nombre de variables par la taille du groupe***. En définitive, il n'est pas nécessaire de prendre des $\phi(x)$ quelconques, ***il nous suffit de garder les moyennes des*** $\phi(x)$ ***par rapport au groupe des symétries***. Ainsi, on n'a besoin de ne garder que les observables "modulo" les symétries. Par exemple, pour les moments d'ordre 2, on ne considère non pas toutes les paires de points possibles, mais on peut les grouper par lots de paires ayant la même distance relative.

Maintenant, comme nous l'avons rappelé K avant de considérer les symétries de $p(x)$ est de l'ordre de e^d. Si l'on veut être capable de le réduire à $K \sim d$, il faut un groupe dont la taille est de l'ordre de e^d. C'est là où on arrive aux limites de l'approche de l'introduction des symétries globales pour réduire le nombre de ϕ. En pratique, on peut concevoir que $p(x)$ est symétrique par tel ou tel groupe (translation, rotation, dilatation), mais ***en définitive on ne dispose pas de suffisamment de connaissances pour réduire drastiquement la dimensionalité du problème***. Il faut donc approcher le problème plus finement. Et en particulier, regarder ***la dépendance du problème à travers les échelles***. En effet, il n'y a pas que des symétries globales qui comptent, ***si l'on observe un phénomène localement, il a tendance à se propager à travers les échelles***. Rendre ce constat en

termes mathématiques nous occupera la prochaine séance.

8.8 NDJE. Transformation de Legendre: cas non convexe

Il s'agit d'un petit complément concernant la transformation de Legendre. La question est que se passe-t-il quand la fonction f n'est pas convexe. Un exemple avec la fonction $f(x) = 1/4 - x^2 + x^4$ en forme de chapeau mexicain (Fig. 29). Si l'on étudie la fonction $g(s) = xs - f(x)$, on doit recourir à la résolution d'une équation du 3eme degré afin d'obtenir le maximum, ce qui ne se peut analytiquement en l'état. Par contre, on sait que $s = f'(x)$ donne la fonction $s(x)$ qui permet de connaitre le supremum. On peut établir le graphe paramétrique en s donné par $\{s = f'(x), g(s(x))\}$ (voir Fig. 29 de droite en bleu). Ce graphe présente une "queue d'hirondelle" caractéristique. La transformée de Legendre $L[f](s)$ est constituée par la courbe en rouge pointillé. Procédons à présent à la transformée de $L[f](t)$, le point important est que la pente de $L[f](s)$ à l'origine est donnée par $\pm x_0 s$ avec $f(\pm x_0) = 0$. Or, comme $L[f](s)$ est convexe, cela signifie que pour tout $0 < x < x_0$, $xs - L[f](s) \leq 0$ et le supremum selon s est atteint en $s = 0$ soit 0. D'où, pour $0 < x < x_0$ $L[L[f]](x) = 0$, et symétriquement elle est aussi nulle pour $-x_0 < x < 0$. En dehors, de l'intervalle $[-x_0, x_0]$, la fonction $f(x)$ est retrouvée, et l'on obtient à présent une courbe manifestement convexe (courbe en rouge pointillé Fig. 29 de gauche). On a alors convexifié la fonction originalement non-convexe.

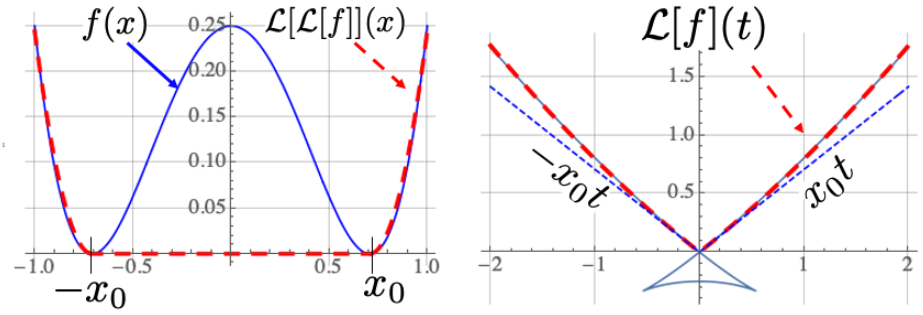


FIGURE 29 – Convexification du chapeau mexicain.

8.9 NDJE. Metropolis-Hasting

C'est un ajout au cours car S. Mallat n'a pu le traiter faute de temps. Il s'agit de l'algorithme de Metropolis-Hastings de génération de chaines de Markov pour obtenir des échantillons selon une distribution cible $P(x)$ pour laquelle on ne dispose pas de méthode directe d'échantillonnage ou que les autres méthodes (ex. Importance samplng) sont peu ou pas efficaces.

La méthode développée d'abord en 1949 par Nicholas C. Metropolis (1915-99) et Stanisław Ulam (1909-84) fut reprise plus en détails en 1953 par Metropolis et collaborateurs puis étendue en 1970 par Wilfred Hastings (1930-2016). On la baptise l'algorithme de Metropolis-Hastings (MH) même si plusieurs auteurs y ont contribué.

Soit donc une distribution $P(x)$ pour laquelle on désire tirer des échantillons. On va construire une chaine de Markov dont on va préciser la matrice de transition $P_{x,x'}$ (rappel passage de x à x') à partir d'une distribution $g(x, x')$ (*proposal*) qui permet de tirer un nouvel échantillon x' connaissant x, et une fonction $A(x, x')$ qui accepte où non ce nouvel échantillon pour fournir un nouvel élément de la chaine, sinon la chaine est complétée avec l'échantillon x (nb. il peut se faire que ce dernier soit donc repris plusieurs fois de suite). Prenons comme expressions de $P_{x,x'}$ et $A(x, x')$:

$$P_{x,x'} = g(x, x')A(x, x') \qquad A(x, x') = \min\left(1, \frac{P(x')g(x', x)}{P(x)g(x, x')}\right) \tag{281}$$

On montre que $P(x)$ ***satisfait l'expression de la balance détaillée*** (Eq. 188). En effet, supposons $P(x')g(x', x) > P(x)g(x, x')$, alors

$$A(x, x') = 1 \qquad A(x', x) = \frac{P(x)g(x, x')}{P(x')g(x', x)} \tag{282}$$

donc

$$P_{x,x'} = g(x, x') \qquad P_{x',x} = g(x', x) \times \frac{P(x)g(x, x')}{P(x')g(x', x)} = \frac{P(x)g(x, x')}{P(x')} \tag{283}$$

d'où

$$P(x)P_{x,x'} = P(x')P_{x',x} \tag{284}$$

ce qui est l'expression désirée. Le cas où $P(x')g(x',x) \leq P(x)g(x,x')$ se traite de la même façon et aboutit à la même conclusion.

Donc, $P(x)$ satisfaisant l'expression de la balance détaillée alors dans les conditions d'ergodicité qui garantissent l'unicité de la mesure invariante, alors $P(x)$ ***est la dite mesure invariante atteinte asymptotiquement***.

L'algorithme se déroule alors ainsi: on génère un nouvel échantillon de la chaine x_i à l'étape i selon la démarche suivante:

1. on tire un échantillon $\hat{x}$ selon $g(x; x_{i-1})$, se peut être une distribution gaussienne de moyenne x_{i-1} par exemple;
2. puis on forme le rapport

$$r = \frac{\tilde{P}(\hat{x})}{\tilde{P}(x_{i-1})} \times \frac{g(\hat{x}, x_{i-1})}{g(x_{i-1}, \hat{x})} \qquad \text{(Metropolis} - \text{Hastings)} \qquad (285)$$

   On utilise la notation $\tilde{P}$ au lieu de P pour signifier que l'on n'a pas besoin en pratique de connaitre la constante de normalisation de la distribution cible.

3. enfin on décide de la valeur de x_i selon:

$$\text{Soit } u \sim \mathcal{U}(0,1), \quad \text{si } r > u \Rightarrow x_i = \hat{x}, \text{ sinon } x_i = x_{i-1} \qquad (286)$$

L'initialisation du processus se fait par la génération aléatoire de x_0. Si la distribution $g(x,y)$ est invariante par échange des variables x et y (ex. une gaussienne) alors le rapport r se simplifie

$$r = \frac{\tilde{P}(\hat{x})}{\tilde{P}(x_{i-1})} \qquad \text{(Metropolis)} \qquad (287)$$

Même si cette version du rapport r est plus particulièrement dédiée à Metropolis, bien souvent on utilise la terminologie Metropolis-Hastings dans tous les cas. L'inconvénient de la méthode est son caractère de marche aléatoire (*random walk*) liée en particulier au fait que la loi $g(x,y)$ est prise une fois pour toute. D'autres méthodes adaptent la forme de cette distribution au fur et à mesure. Enfin, une catégorie de méthodes "Hamiltoniennes", sont des méthodes de Metropolis, qui utilisent l'information de gradient (p variable conjuguée de x) pour réduire le comportement de marche aléatoire. Ces méthodes

sont particulièrement efficaces en grande dimension et dans les cas où la densité de probabilité a une forme complexe. Dans les cas où densité de probabilité présente plusieurs modes disjoints alors il faut procéder autrement encore.

9. Séance du 8 Mars

Durant cette séance, nous allons utiliser l'ensemble des outils développés dans les séances (et cours) précédents afin *de modéliser des processus stochastiques ergodiques non gaussiens* comme des sons, des images, et donc nous allons reprendre les modèles d'***entropie maximum***.

9.1 Modèles d'entropie maximum

On se place dans le cadre où, à partir des données x, on peut calculer un certain nombre de moyennes de fonctions éventuellement (et presque surement) non linéaires $(\mu_k = \mathbb{E}(\phi_k(x)))_{k\leq K}$, et delà on peut inférer une distribution de probabilité d'entropie maximum contrainte par ces observations. Cette distribution paramétrée a pour expression (Th. 17)

$$p_\Theta(x) = Z_\Theta^{-1} \exp\left(-\sum_{k=1}^{K} \theta_k \phi_k(x)\right) \tag{288}$$

avec $\Theta = (\theta_k)_k$ des multiplicateurs de Lagrange associés aux contraintes $(\mu_k)_k$. Nous avons vu que

$$D_{KL}(p\|p_\Theta) = \mathbb{H}(p_\Theta) - \mathbb{H}(p) \geq 0 \tag{289}$$

Il faut donc maximiser l'entropie de p_Θ tout en minimisant $D_{KL}(p\|p_\Theta)$. C'est donc ***le choix des*** $\phi_k(x)$ ***ou leur construction qui nous importe*** maintenant.

Plusieurs problématiques se posent à nous:

- il nous faut calculer des espérances (μ_k) ce qui met en lien la question d'***ergodicité*** (Sec. 6.2). Nous utilisons une information *a priori* sur le processus stochastique étudié à savoir sa ***stationnarité***;
- dans la listes des moments μ_k, nous allons en premier lieu regarder les ***moyennes*** et les ***covariances***. Une des conséquences importantes dans l'*hypothèse stationnaire*

consiste à ce que l'***opérateur de covariance est diagonalisable dans une base de Fourier***, ce qui nous amène à la notion de ***puissance spectrale***. Cependant, nous verrons que la base de Fourier n'est pas appropriée dans la majorité des cas.

- on se place dans le cas ***d'une unique observation*** (ex. 1 seule image) comme en Cosmologie (voir le séminaire d'E. Allys). D'une manière générale, les phénomènes sont ***multi-échelles***. Ainsi, il nous faut procéder à l'***analyse multi-échelles*** (aussi appelée multi-résolutions) qui nous amène à la ***transformée en Ondelettes***. *NDJE. S. Mallat nous donnera un rapide aperçu de cette transformation, pour plus de détails voir Sec. 5.3 Cours 2021 et les références incluses ainsi que le livre cité en avant-propos.*
- Le but de ces notions est d'étudier ***les interactions entre les échelles***. Nous verrons que la capture des structures (ex. filaments des textures pour les images, les impulsions dans les trames sonores, etc) nécessite de ***comprendre comment l'information se propage à travers les échelles***, contrairement au processus gaussiens où les différentes échelles certes sont présentes mais sont indépendantes les unes des autres (elles fluctuent sans interaction). Ces interactions sont capturées avec des ***non-linéarités***, notamment dans ***la transformée de scattering*** (Cours 2020 Sec. 9.5). Nous verrons quelques exemples de générations de sons et d'images et les limites de ce genre d'approche.

9.2 Calcul des moyennes

Nous avons donc 1 échantillon unique x d'une variable aléatoire X (ex. 1 image, 1 trame sonore) et u représente la variable sous-jacente, comme la localisation d'un pixel dans l'image ($u = (u_1, u_2)$). Cette variable u est de basse dimension. Dans l'hypothèse de stationnarité, c'est-à-dire l'invariance par translation sur la variable sous-jacente, les moyennes et les covariances [95] satisfont les relations suivantes:

$$\begin{cases} \mathbb{E}(X(u)) &= \mu(u) = \mu \\ Cov(X(u), X(u')) &= C(u, u') = C(u' - u) \end{cases} \quad \text{(stationnaire)} \qquad (290)$$

95. $Cov(A, B) = \mathbb{E}((A - \mathbb{E}(A))(B - \mathbb{E}(B))^T)$. Bien entendu dans le cas d'images, sons etc, de tailles finies il faut considérer conditions ad hoc pour les bords.

Rappelons que d'après ***l'hypothèse de stationnarité est la manifestation de l'invariance par translation*** de $p(x)$ (Sec. 8.7) la densité de probabilité de X qui sert au calcul des espérances. Autrement dit, ***les espérances ne dépendent pas de la position, et les covariances ne dépendent que des positions relatives***. Dans ce cadre, l'estimation de la moyenne du signal s'obtient par la moyenne empirique

$$\tilde{\mu} = \frac{1}{N} \sum_{u=1}^{N} X(u) \tag{291}$$

Tout d'abord, $\tilde{\mu}$ est un estimateur non biaisé car $\mathbb{E}(\tilde{\mu}) = \mu$ (grâce à l'invariance de l'espérance de $X(u)$). La question qui se pose est la *consistance* de cet estimateur (Cours 2022 Sec. 3.4) c'est-à-dire comment se comporte-t-il quand $N \to \infty$ (ex. le nombre de pixels tend vers l'infini).

Propriété 8

$$\mathbb{E}[(\tilde{\mu} - \mu)^2] = \frac{1}{N} \sum_{\tau=-N+1}^{N-1} \left(1 - \frac{|\tau|}{N}\right) C(\tau) \tag{292}$$

Notons que si $\sum_{\tau=0}^{\infty} |C(\tau)| < \infty$, signifiant que la corrélation entre 2 points diminue rapidement avec l'écart τ entre ces points, alors $\mathbb{E}[(\tilde{\mu} - \mu)^2] = O(1/N)$. L'ergodicité (au moins pour la moyenne) nous fournit l'hypothèse pour que cela fonctionne.

Démonstration 8. La démonstration se fait en écrivant ce que vaut $\tilde{\mu}$ et en procédant à un changement de variable idoine:

$$\begin{aligned}
\mathbb{E}[(\tilde{\mu} - \mu)^2] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{u=1}^{N} X(u) - \mu\right)\left(\frac{1}{N}\sum_{u'=1}^{N} X(u') - \mu\right)\right] \\
&= \mathbb{E}\left[\frac{1}{N^2}\sum_{u,u'=1}^{N} (X(u) - \mu)(X(u') - \mu)\right] \\
&= \frac{1}{N^2}\sum_{u,u'=1}^{N} Cov(u - u') \\
&= \frac{1}{N^2}\sum_{\tau=-(N-1)}^{N-1} \underbrace{(N - |\tau|)}_{\#\text{ termes}} C(\tau)
\end{aligned} \tag{293}$$

Le nombre de termes est calculé en tenant compte que u et u' sont dans $[\![1, N]\!]$. En faisant rentrer un des $1/N$ permet d'obtenir le résultat. ■

Maintenant, la formule nous dit la chose suivante: ***pour que la moyenne empirique converge vers l'espérance, il faut qu'il y ait de la décorrélation***, c'est-à-dire que les $C(\tau)$ décroissent rapidement pour faire converger $\sum_\tau |C(\tau)|$. Dans ce cas $\mathbb{E}[(\tilde{\mu}-\mu)^2] = O(1/N)$.

Donc, ***il faut construire des*** $\phi_k(x)$ ***qui ont de la décorrelation pour assurer l'estimation des moyennes avec 1 échantillon*** (1 image, 1 trame sonore, etc). Cependant, n'utiliser que des moyennes n'est pas suffisant, il nous faut envisager des moments d'ordres supérieurs.

9.3 Moments d'ordre 2 (covariance), échec de Fourier

La matrice de covariance $C = [C(u, u')]_{u,u'}$ (de taille $N \times N$) nous permet de calculer les corrélations entre variables obtenues par combinaison linéaire de X. En effet, $\forall A, B$

$$A = \sum_{u=1}^{N} a(u)X(u) = a^T X,\ \ B = b^T X \Longrightarrow Cov(A, B) = a^T C b \tag{294}$$

(nb. en se rappelant ce que $Cov(A, B)$ (note 95), on obtient facilement le résultat).

La matrice C est ***symétrique***, ***positive*** car $Cov(A, A) = Var(A) = a^T C a \geq 0$ donc tous les vecteurs propres sont associés à des valeurs propres positives (ou nulles). Elle est ***diagonalisable***, et se placer dans la base des vecteurs propres, c'est réaliser une PCA (***Principal Components Analysis*** ou *Analyse en composantes principales*, voir Cours 2021, Sec. 4.3 Th 7). Cependant, on se place dans une hypothèse de stationnarité (Eqs. 290),

donc C est une matrice symétrique et de Toeplitz [96] que l'on peut écrire

$$C = \begin{pmatrix} C(0) & C(1) & C(2) & \dots & C(N-1) \\ C(1) & C(0) & C(1) & \dots & C(N-2) \\ C(2) & C(1) & C(0) & \dots & C(N-3) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ C(N-1) & C(N-2) & \dots & C(1) & C(0) \end{pmatrix} \tag{295}$$

Ainsi, $\forall b$

$$C(u,.)b = \sum_{u'} C(u,u')b(u') = \sum_{u'} C(u-u')b(u') = (C*b)(u) = \sum_{u'} C(u')b(u-u') \tag{296}$$

La présence de ***la convolution*** nous dit que ***la base de diagonalisation est celle de Fourier***. Ainsi

$$C * e^{i\omega u} = \sum_{u'} C(u')e^{i\omega(u-u')} = e^{i\omega u} \underbrace{\sum_{u'} C(u')e^{-i\omega u'}}_{\widehat{C}(\omega)} \tag{297}$$

où l'on reconnait ***la transformée (série -) de Fourier de l'opérateur de covariance et les sinusoïde sont les v.p de cet opérateur***. Comme *l'opérateur est positif* donc

$$\widehat{C}(\omega) \geq 0 \qquad \text{(puissance spectrale)} \tag{298}$$

Notons que

$$C(u) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widehat{C}(\omega) e^{i\omega u} \, d\omega \tag{299}$$

et en particulier, en $\tau = 0$,

$$C(0) = \mathbb{E}[|X(u)|^2] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widehat{C}(\omega) \, d\omega \tag{300}$$

$C(0)$ est l'énergie totale du signal, et c'est pour cette raison que *la puissance spectrale peut être assimilée à la densité d'énergie par fréquence.*

96. Otto Toeplitz (1881-1940).

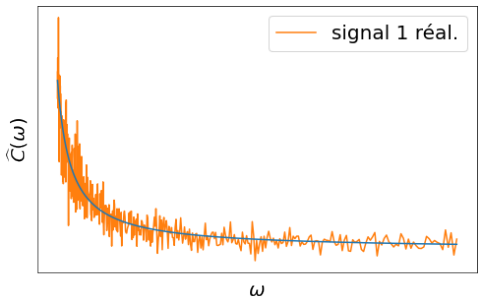


FIGURE 30 – Schématisation de l'évolution de la puissance spectrale en fonction de ω (bleu). Pour des images $\widehat{C}(\omega) \sim 1/\omega$. Lorsque l'on a 1 unique réalisation alors ce que l'on observe c'est une auto-corrélation du signal (orange) dont les fluctuations sont gouvernées également par $\widehat{C}(\omega)$.

Comment estimer cette covariance, si l'on a qu'une seule réalisation X? On peut projeter [97] X sur un vecteur propre $e_\omega(u) = e^{i\omega u}/\sqrt{N}$:

$$e_\omega^T\ X = \frac{1}{\sqrt{N}} \sum_{u=1}^{N} X(u) e^{-i\omega u} \tag{301}$$

D'après la propriété (Eq. 294), comme $a = (e_\omega(u))_{1 \leq u \leq N}$ est un *vect.p* de C de *va.p* $\widehat{C}(\omega)$, on en déduit que

$$\mathbb{E}(|e_\omega^T\ X|^2) = \widehat{C}(\omega) \tag{302}$$

Une évolution typique de $\widehat{C}(\omega)$ est donnée sur la figure 30 (courbe bleue).

Cependant, dans le cas d'*une seule réalisation*, l'auto-corrélation du signal fluctue (courbe orange), et l'amplitude de ces fluctuations est directement reliée à la valeur de $\widehat{C}(\omega)$. On a donc des erreurs et la question se pose de ***comment obtenir un estimateur consistant***? Pour ce faire, on choisit de moyenner l'auto-correlation sur des plages [98] de valeurs de ω_k pondérées par des fenêtres bien choisies, afin d'estimer la covariance uniquement en quelques valeurs de ω. Ainsi, en combinant plusieurs mesures de $\widehat{C}(\omega_k)$ le bruit peut être combattu. Mais, la consistance n'est réalisée que pour quelques valeurs de ω_k. ***On a un problème intrinsèque de consistance de l'estimateur de la covariance avec la***

97. NDJE. dans le cas complexe: $a^T b$ doit être compris $(a^*)^T\ b$, où $*$ signifie le conjugué complexe.
98. NDJE. k indique que l'on prend des valeurs discrètes de ω en pratique quand on numérise le signal

transformée de Fourier dans les cas pratiques.

Il y a une raison plus profonde pour laquelle la transformée de Fourier n'est pas adaptée. Quand on analyse une trame sonore avec des transitoires, ou bien une image avec des contours d'objets, ce qu'il se produit est la chose suivante: quand on effectue la transformée de Fourier d'un signal tel que celui présenté sur la figure 31 (image du haut, courbe bleue), on le corrèle avec des sinusoïdes de type de la courbe orange, c'est-à-dire que l'on procède à un produit scalaire qui mélange l'information du signal sur tout son déroulé "temporel" (nb. le phénomène est identique en 2D pour une image). ***En particulier, à une fréquence donnée, on mélange sans distinction les temps où le signal est régulier, et les temps où au contraire il présente beaucoup de structures***. Quand on essaye de reconstruire alors le signal, en ne prenant qu'une partie de ses composantes spectrales (par ex. ω tq. $|\widehat{C}(\omega)| > T$, où T est un seuil) alors il apparait des petites oscillations à tous les moments de transitions (courbe rouge, image du bas). En fait, ***la dépendance du spectre de Fourier du signal est dominé par les transitoires du signal***: à la lecture du spectre, on ne sait pas dire que le signal est régulier en telle ou telle partie de son déroulé temporel (ou partie d'une image).

9.4 Filtrage par Ondelettes (1D)

Les transitoires sont typiques des phénomènes non-gaussiens, pensez par exemple à la localisation des tourbillons dans un flot turbulent. Donc, il nous importe de savoir les traiter le plus fidèlement possible. Il faudrait pouvoir avoir ***une description locale du signal***, et donc pouvoir ***le corréler plutôt avec des fonctions qui oscillent uniquement sur une petite fenêtre "temporelle"*** du type de celle en vert sur l'image du haut de la figure 31, que l'on nomme ***ondelette***. Afin de capturer les transitoires sur l'ensemble de la plage "temporelle" du signal, il faut pouvoir ***translater*** ces ondelettes, et afin de capturer toutes les échelles de temps il faut également les ***dilater/contracter*** comme sur la figure 32. L'intérêt de ces ondelettes est que pour l'une d'entre elles, ***le coefficient de la corrélation avec le signal est non nul que si le signal présente un transitoire localisé dans le support de l'ondelette***.

Si $\psi_\lambda(u)$ est la fonction "ondelette" dont le support est d'échelle λ, on calcule alors

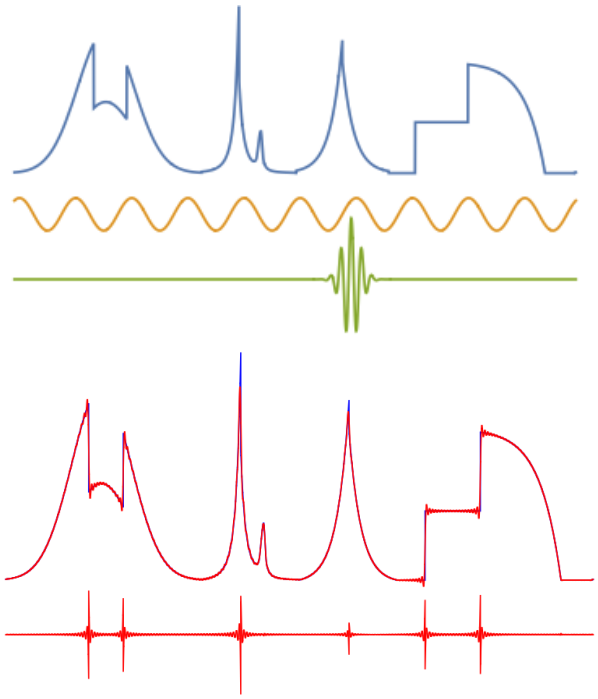


FIGURE 31 – Image du haut: signal original en bleu, sinusoïde délocalisée typique d'une analyse par transformée de Fourier (orange) et ondelette localisée typique d'une analyse par transformée en Ondelettes (vert). Image du bas: signal reconstruit en rouge quand on ne prend qu'une partie du spectre de Fourier (la différence avec le vrai signal est donnée sur le graphe en bas). Voir *gibbs_FFT.ipynb*

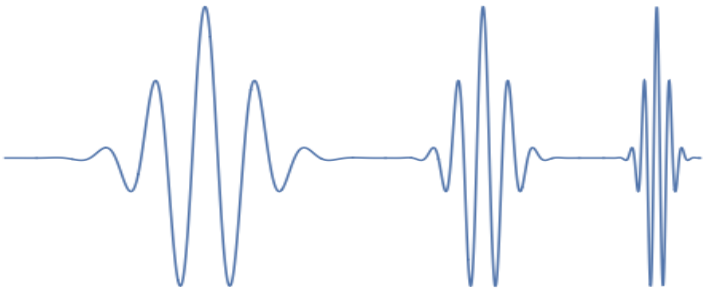


FIGURE 32 – Illustration des opérations de translation et changement d'échelle appliquée à une ondelette.

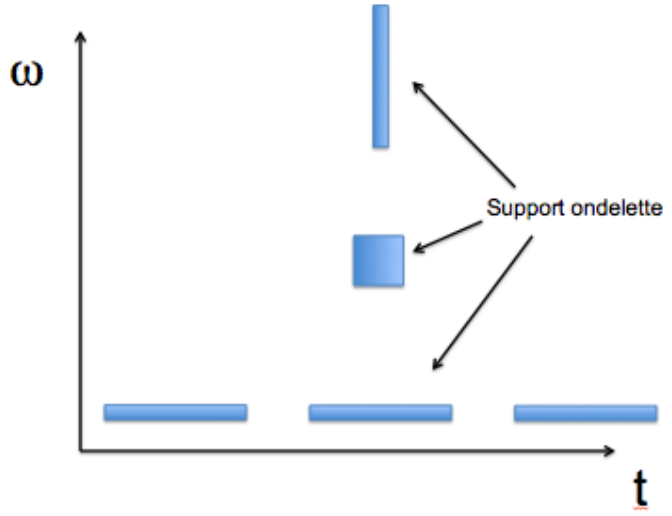


FIGURE 33 – Changement de la taille du support de l'ondelette suivant la fréquence, et par contre invariance suivant une translation en temps.

le coefficient d'ondelette [99] de $X(u)$ selon

$$(X * \psi_\lambda)(u) = \sum_{u'} X(u')\psi_\lambda(u - u') = (W_\lambda x)(u) \tag{303}$$

Ce coefficient mesure la fluctuation du signal au voisinage de u ***sur une échelle de l'ordre de*** λ. En quoi cela nous résout le problème de la transformée de Fourier et des fluctuations de $\widehat{C}(\omega)$? Le point principal est que ***le support de l'ondelette est borné à la fois en temps et en fréquence*** [100] comme cela est illustré sur la figure 33. On réalise alors automatiquement le fenêtrage mentionné au paragraphe précédent afin d'obtenir des estimateurs des moments d'ordre 2 qui soient consistants.

Soit l'ondelette de base la fonction suivante, dite ondelette de Morlet ou de Gabor [101]

99. NDJE. Pour une introduction détaillée de la transformée en Ondelettes voir par exemple le Cours de 2020 Sec 7. et le Cours 2021 Secs. 7 et 8.
100. NDJE. On a également le même phénomène de concentration pour des ondelettes 2D.
101. NDJE. de Jean Morlet (1931-2007) et Dennis Gabor (1900-1979).

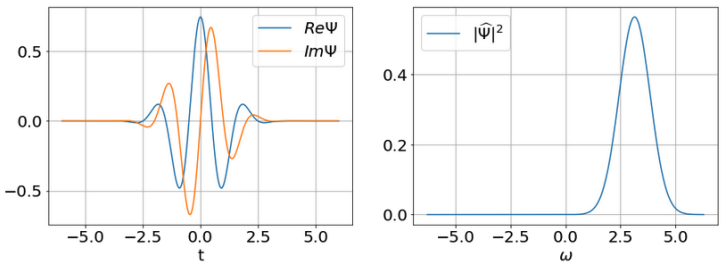


FIGURE 34 – Ondelette complexe de Morlet Eq. 304 dont le spectre de puissance est quasi nul pour $\omega < 0$ ($\sigma = \pi$).

(Voir notebook *morlet_wave_1D_2D.ipynb*)

$$\psi(x) = \frac{e^{-\frac{t^2}{2}}\left(-e^{-\frac{\sigma^2}{2}} + e^{i\sigma t}\right)}{\sqrt[4]{\pi}\sqrt{-e^{-\sigma^2} - 2e^{-\frac{3\sigma^2}{4}} + 1}} \qquad \widehat{\psi}(\omega) = \frac{e^{-\frac{1}{2}(\sigma-\omega)^2} - e^{-\frac{\sigma^2}{2}-\frac{\omega^2}{2}}}{\sqrt[4]{\pi}\sqrt{-e^{-\sigma^2} - 2e^{-\frac{3\sigma^2}{4}} + 1}} \tag{304}$$

Elle a la particularité d'avoir un spectre de Fourier quasi nul pour $\omega \leq 0$ (ondelette admissible) et concentré autour de la fréquence ω_0 solution de

$$\omega_0 = \frac{\sigma}{1 - e^{-\sigma\omega_0}}$$

($\sigma \gtrsim 2$ alors $\omega_0 \sim \sigma$). La figure 34 illustre le propos pour $\sigma = \pi$.

Notons que l'on peut grosso modo construire une ondelette de base à partir d'une fenêtre (ex. gaussienne) multipliée par une phase afin que le spectre de puissance soit shifté de la valeur souhaitée dans le domaine $\omega > 0$. Le fait que $\widehat{\psi}(0) = 0$ signifie que la moyenne de l'ondelette est nulle:

$$\widehat{\psi}(0) = 0 \Rightarrow \sum_u \psi(u) = 0 \tag{305}$$

Maintenant, si nous appliquons un facteur d'échelle λ, il vient alors

$$\psi_\lambda(u) = \lambda\psi(\lambda u) \Longrightarrow \widehat{\psi_\lambda}(\omega) = \widehat{\psi}(\lambda^{-1}\omega) \tag{306}$$

En jouant sur les valeurs de λ on peut couvrir toute la gamme de ω (Fig. 35 haut). Notons

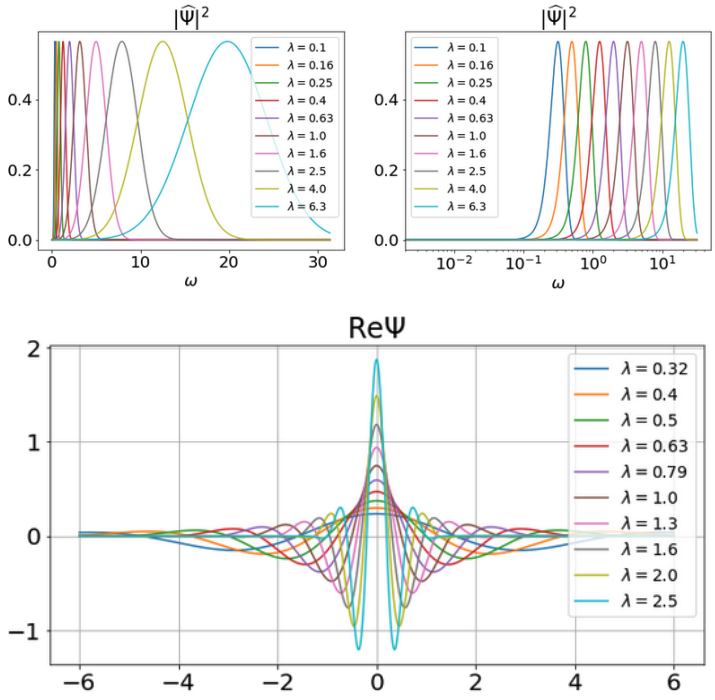


FIGURE 35 – Image du haut: Spectre de Fourier de l'ondelette de Morlet (Fig. 34) pour différents facteurs d'échelle (échelle linéaire à gauche, logarithmique à droite). Remarquez la constance de la largeur du spectre en échelle logarithmique. Image du bas: Ondelette de Morlet dans l'espace réel pour différentes valeurs du facteur d'échelle. Remarquez que la largeur caractéristique de l'ondelette dans l'espace réel évolue en sens inverse de sa largeur en Fourier.

en passant que le spectre de Fourier d'une ondelette de Morlet a une largeur constante en échelle logarithmique. Remarquons une caractéristique générale que ***plus le spectre en Fourier est large, plus la largeur caractéristique de l'ondelette dans l'espace réel est étroit, et vice-versa*** (Fig. 35 bas).

Lorsque l'on effectue une convolution du signal X par une ondelette ψ_λ dans l'espace réel, dans l'espace de Fourier pour mémoire on effectue le produit des transformées de Fourier comme suit

$$\widehat{W_\lambda x}(\omega) = \widehat{X * \psi_\lambda}(\omega) = \widehat{X}(\omega) \times \widehat{\psi_\lambda}(\omega) \tag{307}$$

Donc, chaque ondelette va capturer une partie du spectre de Fourier du signal suivant la

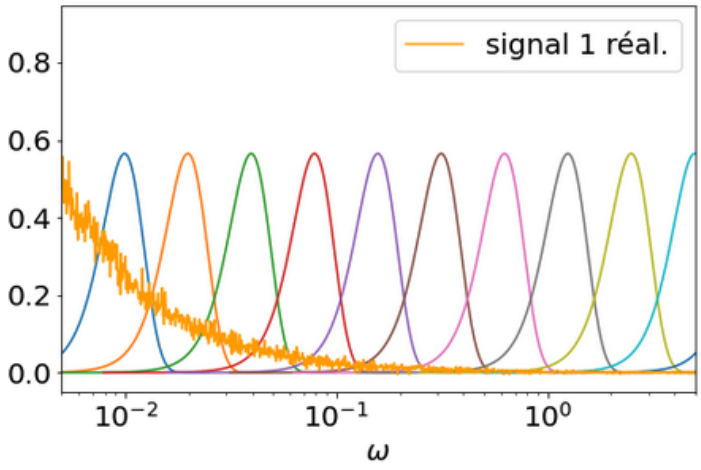


FIGURE 36 – Evolution de $\widehat{C}(\omega)$ (Fig. 30) et spectres de quelques ondelettes obtenues avec différents facteurs d'échelle.

valeur de λ, et en même temps elle capture localement l'évolution du signal dans l'espace réel.

On va appliquer ce type de filtrage dans le cas de $\widehat{C}(\omega)$ comme cela est illustré sur la figure 36. ***Non seulement cela donne des estimateurs de la puissance spectrale qui sont consistants, mais dans le cas non-gaussien cela nous donne un renseignement utile sur les transitoires***.

NDJE. Avant de passer au cas 2D, faisons un petit ajout que S. Mallat n'a pu couvrir faute de temps que vous pouvez retrouver par exemple dans les Sec. 7.5 et 8.1 du Cours de 2020. Quand on regarde l'évolution des spectres de Fourier des ondelettes (Fig. 35), pour couvrir l'ensemble des valeurs $\omega \in [0, \infty]$ *il faudrait une infinité d'ondelettes. Considérant le cas des hautes fréquences, on sait que la puissance spectrale du signal diminue, donc on peut imaginer de ne pas prendre en compte les ondelettes dont le facteur d'échelle est au-delà d'un cut-off* $\lambda \geq \lambda_{max}$*. Par contre, à basse fréquence, on ne peut en faire de même car on sait que le signal a des plages de régularité (pensez aux zones de même intensité dans une image quelconque). On peut néanmoins compléter par exemple la famille de filtres passe-bande* $\{\psi_\lambda\}_{\lambda_{min} \leq \lambda \leq \lambda_{max}}$ *par un filtre passe-bas (ex. une gaussienne)* ϕ*. Alors le spectre de Fourier ainsi couvert permet de reconstruire le signal à partir des coefficients*

d'ondelettes et de l'équivalent obtenu avec le filtre ϕ, c'est-à-dire

$$Wx = (\psi_\lambda * X), \phi * X) \tag{308}$$

Enfin, suivant le schéma des "largeurs/hauteurs" caractéristiques de la puissance de l'ondelette dans le plan temps-fréquence (Fig. 33), on peut montrer que l'on peut discrétiser à la fois les échelles et les translations spatiales selon $(2^j\ n, 2^j)$ avec $(n, j) \in \mathbb{Z}^2$, pour obtenir une base orthonormale de $L^2(\mathbb{R})$. Dans l'audio, en pratique il est d'usage d'utiliser un peu de redondance en prenant un facteur d'échelle $2^{j/Q}$ pour les ondelettes avec $Q \sim 16$. Ceci est en lien avec l'audition humaine et la notion d'octave (Voir. Cours 2020 Sec. 8.1).

9.5 Filtrage en 2D

NDJE. Pour le cas 2D traité en détails voir le Cours 2021 Sec. 8.4.

Le principe général est le suivant: ce que l'on peut mettre en place en 1D, peut être étendu à nD quelconque. En Fourier si $X(u = (u_1, u_2))$ alors si $\omega = (\omega_1, \omega_2)$ et $\omega.u = \sum_{i=1,2} \omega_i u_i$, nous avons synthétiquement

$$\widehat{X}(\omega) = \sum_u X(u) e^{-i\omega.u} \tag{309}$$

Comme pour le cas 1D, il nous faut concevoir une ondelette admissible dont le spectre de puissance soit limité à une zone du plan de Fourier. Une solution possible est de copier la philosophie de l'ondelette de Morlet/Gabor 1D (voir le notebook *morlet_wave_1D_2D.ipynb*):

$$\begin{aligned}
\psi(u_1, u_2) &= \left(-e^{-\frac{\xi^2}{2}} + e^{i\xi u_1}\right) e^{-\frac{u_1^2+u_2^2}{2\sigma^2}} \\
\widehat{\psi}(\omega_1, \omega_2) &= \pi\sigma^2 e^{-\frac{1}{2}\sigma^2\omega_2^2} \left(e^{-\frac{1}{2}\sigma^2(\omega_1-\xi)^2} - e^{-\frac{1}{2}(\xi^2+\sigma^2 w_1^2)}\right)
\end{aligned} \tag{310}$$

Le facteur σ joue sur la largeur caractéristique de l'ondelette tandis que ξ fixe la position du maximum de puissance en Fourier. Une illustration est donnée sur la figure 37. Maintenant, comme en 1D on peut changer l'échelle λ que l'on prend dyadique ($\lambda = 2^j$ avec

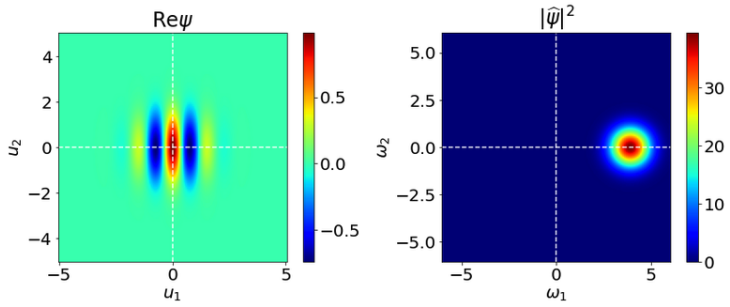


FIGURE 37 – Ondelette de base en 2D (Eq. 310) dont le spectre en Fourier est localisé autour de $\omega_1 \approx \xi$ ($\sigma = 1$, $\xi = 5/4\pi$) et dont la puissance dans le plan $\omega_1 \leq 0$ est nulle.

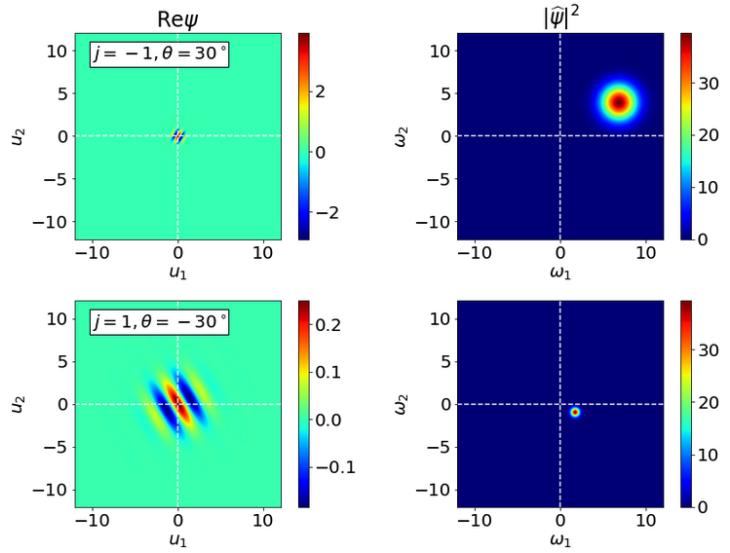


FIGURE 38 – Effet d'un changement d'échelle et d'une rotation sur l'ondelette de base en 2D (Eq. 311): pour la ligne du haut $j = -1$ et $\theta = 30°$ tandis que pour la ligne du bas $j = 1$ et $\theta = -30°$.

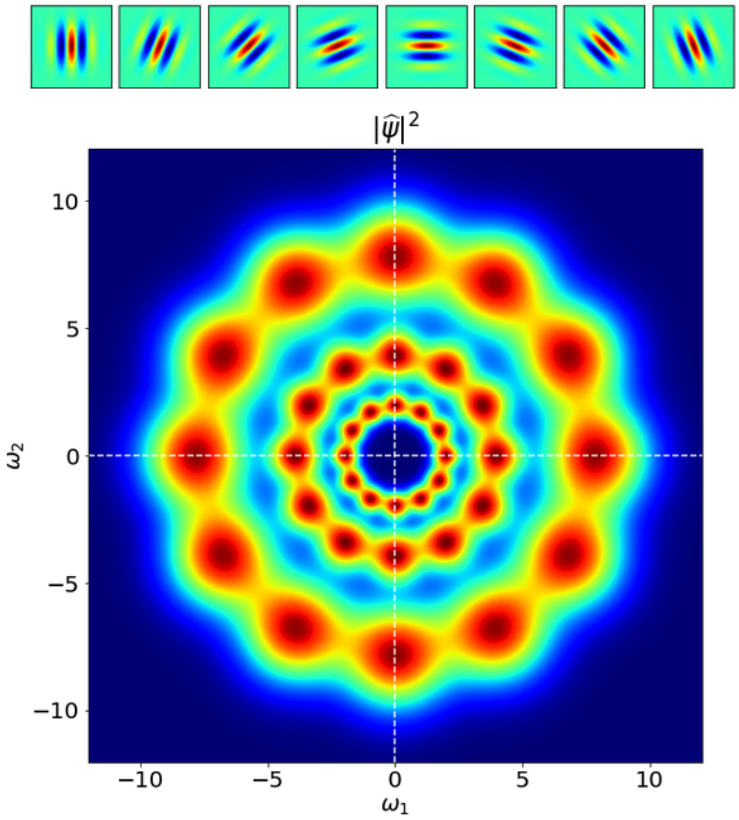


FIGURE 39 – Haut: ondelette (partie réelle) ayant subit des rotations $k\pi/8$ avec $k \in [\![0, 7]\!]$. Bas: couverture d'une partie du plan de Fourier en collectant les spectres de plusieurs ondelettes de facteur d'échelle $j = (-1, 0, 1)$ et ayant subit les rotations multiples de 30°.

$j \in \mathbb{Z}$), mais on peut également appliquer une rotation d'un angle θ. Il vient [102]

$$\psi_{j,\theta}(u) = 2^{-2j}\psi(2^{-j}r_\theta.u) \Longrightarrow \widehat{\psi_{j,\theta}}(\omega) = \widehat{\psi}(2^j r_{-\theta}.\omega) \tag{311}$$

Une illustration de l'action combinée d'un changement d'échelle et d'une rotation est donnée sur la figure 38. Ensuite, on combinant une collection d'ondelettes avec différents paramètres j et θ, on peut couvrir le plan de Fourier comme 1D. La figure 39 illustre le propos. Bien entendu, on peut adapter l'ondelette de base et le nombre de rotations

102. NDJE. *primo* au regard du Cours de 2020 Sec. 8.2, j'ai changé la définition du sens de rotation qui est purement arbitraire. Ici $r_\theta = \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$. Secundo, le facteur de normalisation est bien λ^2 car on a 2 coordonnées.

pour combler les éventuels trous laissés ici juste pour rendre visibles les différents "blobs" associés à chaque ondelettes. Cependant comme en 1D, pour couvrir les basses fréquences (la zone centrale, $\omega \sim (0,0)$), on a recourt à l'usage un filtre passe-bas $\phi(\omega)$ comme une simple gaussienne.

Donc finalement, ***l'image (signal) est passée à travers l'ensemble d'une série de filtres passe-bande et du filtre passe-bas*** ($\theta_k = 2\pi k/K$) tout comme 1D

$$Wx = ((\psi_{j,\theta_k} * X)_{j \in [\![j_{min}, j_{max}]\!], k \in [\![0, K-1]\!]}, \phi * X) \tag{312}$$

ce qui permet d'analyser le signal à toutes les échelles et toutes les orientations. Il faut également tenir compte des translations dans l'espace réel afin d'avoir une vision complète des transitoires dans la totalité de l'image. Notez qu'une orientation particulière d'une ondelette est sensible à des transitoires le long de l'axe perpendiculaire.

9.6 Exemples d'usage

NDJE. S. Mallat nous projette quelques exemples de décompositions en ondelettes, j'essaye de les reproduire ici. Certains sont issus de la section 2.9.2 que je replace ici pour la clarté du propos.

Un exemple de décomposition d'un signal audio est donné sur la figure 40 (Voir aussi des exemples simples dans le notebook *wavelet1D.ipynb*). L'axe horizontal est celui du temps, tandis que l'axe vertical est celui des fréquences. Mais comme la taille des filtres passe-bandes est indépendante de l'échelle en échelle logarithmique (ex. ondelette de Morlet) et que la position du maximum est reliée à l'échelle, donc l'axe vertical est également un repérage en $\log \lambda$. En tout point de ce plan on représente par une couleur le module de Wx avec en bleu la valeur 0. Notez que la quasi-totalité des coefficients est nulle. Seuls ne comptent que quelques coefficients. ***Ainsi, on a une représentation dans le plan temps-(échelle log.) de la position des transitoires présents dans le signal*** (à quel temps, et à quelle échelle λ). En l'occurrence, on peut décrire les différentes "attaques" de notes avec leurs fréquences fondamentales et leurs harmoniques. On voit donc les structures du signal qui vont permettre de discriminer/identifier tel ou tel instrument, etc.

L'exemple suivant est celui de la figure 41 où une note est jouée par plusieurs instruments de musique. L'analyse par ondelettes permet d'avoir une représentation fine des

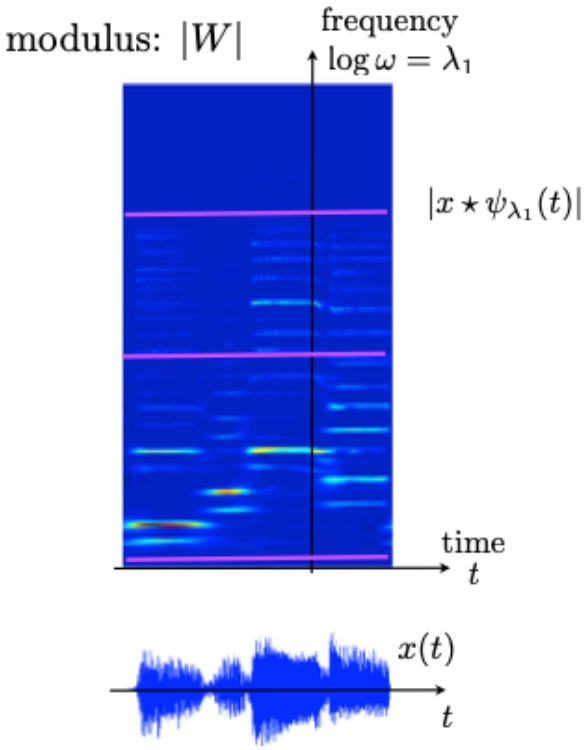


FIGURE 40 – Exemple de décomposition en Ondelettes d'un signal $x(t)$: on représente en couleur la valeur de $\|Wx\|$ avec le bleu représentant des valeurs nulles. On se rend compte de la parcimonie de la décomposition en fréquence et de son évolution temporelle.

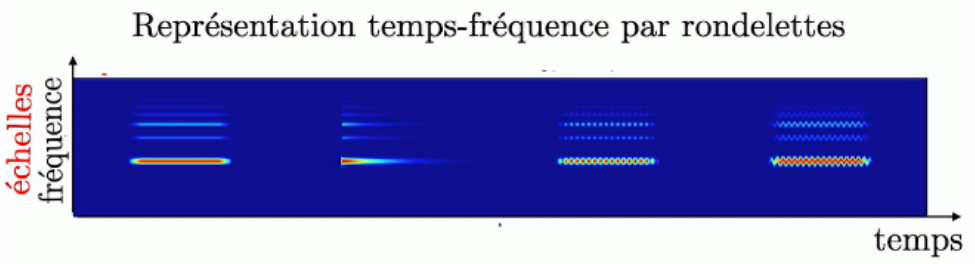


FIGURE 41 – Représentation Temps-Fréquence de quatre séquences de musique (nb. le "r" de rondelettes est une coquille sympathique).

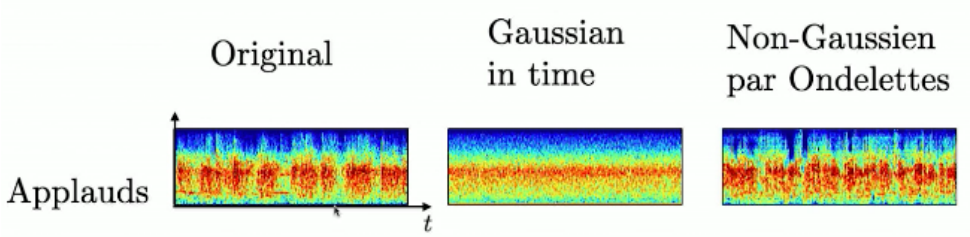


FIGURE 42 – Représentations Temps-Fréquence d'un applaudissement: le son original à gauche; au milieu la génération d'un modèle gaussien qui ne conserve que l'énergie originale à toutes les échelles/fréquences; à droite la génération par un modèle plus réaliste qui tient compte des corrélations entre échelles.

structures, d'où l'intérêt d'avoir ce réseau de nombreux filtres temps-échelle. ***Donc, à partir d'un signal*** X ***qui à*** d ***valeurs temporelles, on va le projeter dans un espace de plus grande dimension ce qui est une méthode générale en Machine Learning***. Cependant, dans le cas présent, seulement quelques coefficients portent de l'information qui sert par la suite à l'analyse (ex. classification...).

Le troisième exemple concerne celui de la figure 42 qui montre 3 représentations d'un "applaudissement". On peut effectuer une décomposition en ondelettes du son original $X(u)$ et la représenter dans le plan temps-échelle comme précédemment. Mais, on peut essayer de reproduire le son en le modélisant par un processus gaussien, c'est-à-dire en capturant les moments du 1er et 2nd ordre du son original à chaque instant (discret). Ainsi sur une tranche en temps on peut modéliser la densité de probabilité $p(x)$ du signal par

$$\tilde{p}(x) \sim \exp\{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\} \tag{313}$$

et effectuer une nouvelle réalisation $X'(u)$ que l'on analyse de nouveau par ondelettes. On remarque que la répartition de la puissance selon les fréquences (échelles) du signal en fonction du temps est bien respectée. Par contre, ***on a perdu les structures du signal qui ne sont pas capturées par la moyenne et la covariance***, ce que l'on comprend à présent car le processus gaussien ne peut en tenir compte. Donc, il va falloir trouver un moyen de capturer les structures manquantes. La troisième représentation sera expliquée plus tard dans la séance.

Il en va de même si l'on considère un phénomène turbulent en 2D (Fig. 43 image de

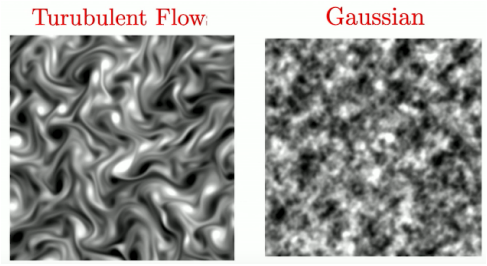


FIGURE 43 – Exemple d'une simulation d'un écoulement turbulent en 2D (gauche), et de la réalisation (droite) d'un modèle gaussien qui simplifie le problème en ne prenant en compte que les moments d'ordre 2 de la distribution de gauche c'est-à-dire en ne considérant que la matrice de covariance.

gauche). A partir de $d = 6\ 10^4$ échantillons du signal X, on peut utiliser $\mu = \mathbb{E}(X)$ et n moments d'ordre 2 $C(k) = \mathbb{E}(\phi_k(X))$ définis par $\phi_k(x) = (k.x - \mu)(x - \mu)$ avec $k.x$ une translation de x par $k \in \mathbb{R}^2$. On peut alors construire un modèle gaussien du type

$$\tilde{p}(x) \sim \exp\left(-\sum_k \beta_k \phi_k(x)\right)$$

Or l'image de droite de la figure 43 nous donne le résultat d'une réalisation à partir de ce modèle gaussien. Force est de constater que si les échelles de fluctuations semblent respectées, ***la géométrie des structures n'est pas celle du signal original***. Donc, le problème qui se pose en physique mais d'une manière générale en traitement su signal, est de savoir capturer les structures du signal au-delà des moments d'ordre 2.

Quand on utilise des ondelettes 2D, avec différents facteurs d'échelles dyadiques, différentes orientations et translation spatiales, on "éclate" une image typique dans tous ces canaux de filtrages et l'on obtient des imagettes que l'on peut représenter comme sur la figure 44. On remarque qu'à chaque échelle, l'orientation des ondelettes rend visible les transitions orthogonales dans l'image d'origine. ***On détecte alors les frontières entre zones d'égale intensité, ce sont les phénomènes transitoires recherchés***. Remarquons que cette façon d'opérer les différents filtres d'ondelettes ressemble à s'y méprendre à ***la structure des réseaux de neurones convolutionnels***.

Ce que l'on constate aussi ici en 2D, et nous l'avions remarqué également en 1D, dès

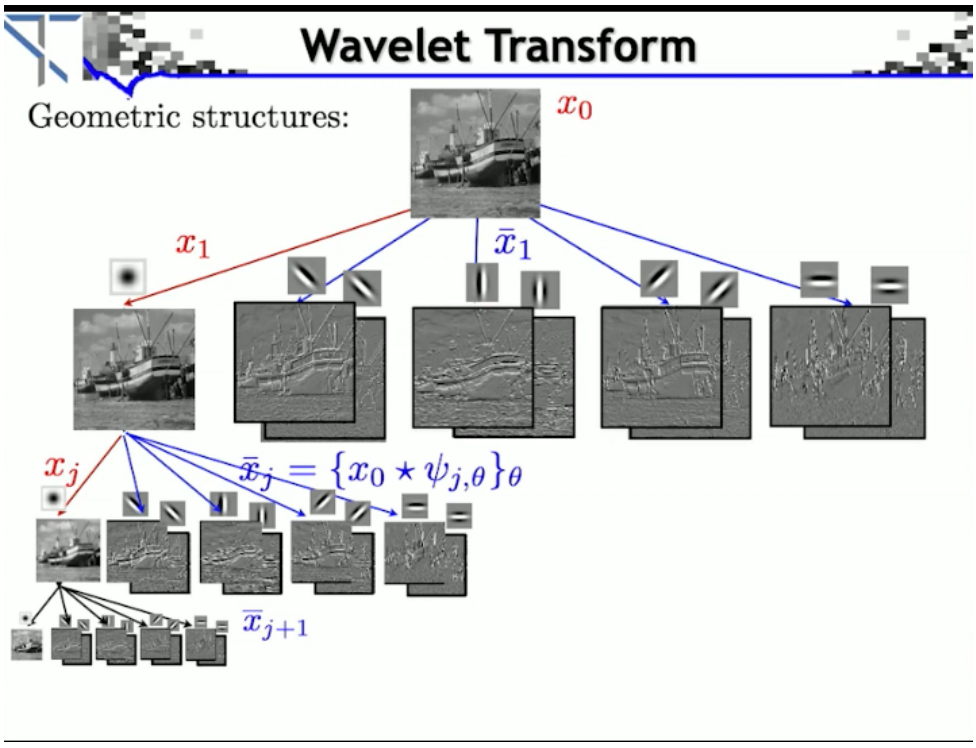


FIGURE 44 – Cascade successive d'applications des filtres passe-bas (ϕ) et passe-bande ψ à différentes échelles et rotations. A une échelle et rotation donnée, on collecte tous les coefficients obtenus par les translations afin de reconstituer des imagettes. A chaque changement d'échelle, on opère un pooling/moyennage d'un facteur 2 dans les deux directions ce qui change la taille des imagettes. Les zones sombres indiques des zones pour lesquelles les coefficients sont nuls. Le notebook *wavelet2D_sparsity.ipynb* permet de calculer la première étape de décomposition avec une ondelette discrète de Haar (Cours 2012 Sec. 8.4)

qu'il y a des transitoires entre plages très régulières, les coefficients d'ondelettes sont essentiellement nuls, sauf pour quelques uns qui sont très localisés dans le plan temps-échelle. C'est ce que l'on appelle ***la parcimonie*** (*sparsity* en anglais). De plus entre les échelles on remarque à l'œil que les structures sont corrélées, il y a ***beaucoup de dépendance entre les échelles***. Comment capturer ces dépendances à l'aide de quantités statistiques?

9.7 La parcimonie

NDJE. Voir également le cours 2021 Sec. 5.1. dans le cas d'une base orthonormée où l'on opère un seuillage sur les coefficients de la décomposition. Ici S. Mallat utilise une autre vision dans le cadre des estimateurs de moyennes.

Supposons que l'on veuille utiliser $\mathbb{E}(|X|)$ et $\mathbb{E}(|X|^2)$, on utilise alors des estimateurs (ici u peut être vu comme l'indice d'un pixel en 2D)

$$\tilde{\mu}_1 = \frac{1}{N}\sum_{u=1}^{N}|X(u)| \qquad\qquad \tilde{\mu}_2 = \frac{1}{N}\sum_{u=1}^{N}|X(u)|^2 \tag{314}$$

Que nous dit le rapport $\tilde{\mu}_1^2/\tilde{\mu}_2$?

Propriété 9

$$\frac{1}{N} \leq r = \frac{\tilde{\mu}_1^2}{\tilde{\mu}_2} \leq 1 \tag{315}$$

Le cas où $r = 1$ correspond à un signal constant, et $r = 1/N$ quand le signal est concentré sur 1 pixel dans le cas d'une image où 1 échantillon temporel en 1D.

La démonstration $r \leq 1$ se fait en utilisant astucieusement Cauchy-Schwartz, $(\sum 1 \times |x(u)|)^2 \leq (\sum 1^2) \times (\sum |x(u)|^2)$. La première inégalité tient au fait que le carré de la somme des valeurs absolues est toujours plus grand que la somme des carrés des valeurs absolues.

Donc, ***la valeur de r nous mesure en quelque sorte le caractère parcimonieux du signal*** X: très parcimonieux quand r est petit ou absolument pas pour $r = 1$. Dans le cas d'un signal gaussien, qu'en est-il? Pour $x \sim \mathcal{N}(0, 1)$ en 1D alors $(\mu_1, \mu_2) = (\sqrt{2/\pi}, 2)$ donc $r_g = 2/\pi$, et en 2D pour $x \sim \mathcal{N}(0, \mathbf{1}_{2\times 2})$ (cas d'une gaussienne complexe aussi) alors

$(\mu_1, \mu_2) = (\sqrt{\pi/2}, 2)$ donc $r_g = \pi/4$. Qu'en est-il d'une image, comme les imagettes de la décomposition Fig. 44? ***Si l'on a*** $r < r_g$ ***alors on a une information de parcimonie qui est bien supérieure au cas gaussien***. *NDJE. Le notebook wavelet2D_sparsity.ipynb nous donne un example et fournit des valeurs de* $r \approx (0.20 - 0.25) < r_g$ *donc illustrant une sparsité des coefficients d'ondelettes bien plus grande qu'un processus gaussien.*

Ces notions de parcimonie ont été beaucoup utilisées dans les années 1980-2000 nous dit S. Mallat. C'est très utile pour faire du codage notamment (Voir Cours 2022 Sec. 9.5 par exemple), mais ce sont des informations qui sont globales. Elles ne sont pas suffisantes pour pouvoir contraindre un modèle permettant de générer une nouvelle image de bateau par exemple.

Ce qu'il faut capturer comme dit à la section précédente se sont les ***corrélations à travers les échelles***.

9.8 Interactions entre les échelles

Comprendre ces interaction entre les différentes échelles, c'est vraiment *le point fondamental* de tout ce domaine des Mathématiques et de la Physique.

En utilisant les différents coefficients d'ondelettes, on peut étudier la corrélation

$$C(u, u', \lambda, \lambda') = Cov((X * \psi_\lambda)(u), (X * \psi_{\lambda'})(u')) \tag{316}$$

Une propriété nous dit que si les supports des transformées de Fourier des 2 ondelettes sont disjoints alors le coefficient de corrélation est nul

$$C(u, u', \lambda, \lambda') = 0 \qquad \text{si} \qquad \widehat{\psi_\lambda)}(\omega) \times \widehat{\psi_{\lambda'}}(\omega) = 0 \tag{317}$$

C'est un résultat qui vient du filtrage de processus. Soit X un processus aléatoire stationnaire, on peut calculer la convolution avec un filtre déterministe h. Or,

$$(g_\tau X)(u) = X(u - \tau) \Longrightarrow g_\tau(X * h) = (g_\tau X) * h \tag{318}$$

qui est la manifestation de l'équivariance de la convolution (Déf. 21). Et comme $g_\tau X = X$ (stationnaire),

$g_\tau(X * h) = X * h$ donc $X * h$ est un processus stationnaire. En particulier

$$\mathbb{E}[(X * h)(u)] = \mathbb{E}\left[\sum_v X(u-v)h(v)\right] = \mathbb{E}[X]\sum_v h(v) \tag{319}$$

Dans le cas où h est une ondelette admissible $\sum_v h(v) = 0$ (Eq. 305), et donc

$$\mathbb{E}[(X * h)(u)] = 0 \tag{320}$$

Concernant les corrélations entre $X * h$ et $X * g$ pris en 2 points différents, alors [103]

$$Cov((X * h)(u), (X * g)(u')) = (C * h * \tilde{g})(u - u') = C_{hg}(u - u') \tag{321}$$

avec $C(\tau) = \mathbb{E}(X(u)X(u-\tau))$ la fonction de corrélation à 2 points dont la transformée de Fourier est le spectre de puissance $\hat{C}(\omega)$ (que l'on mesure par exemple dans les analyses cosmologiques du fond diffus micro-ondes), et $\tilde{g}(u) = g^*(-u)$.

Cela dit que $(X * h)$ et $(X * g)$ sont stationnaires de façon jointe (Eq. 290) et l'on peut calculer la transformée de Fourier, il vient

$$\widehat{C_{hg}}(\omega) = \hat{C}(\omega)\hat{h}(\omega)\hat{g}^*(\omega) \tag{322}$$

Donc, cela nous dit que pour un processus X stationnaire, si l'on veut estimer la corrélation entre deux coefficients d'ondelettes admissibles, cela revient à filtrer la covariance du processus par les deux filtres correspondants. D'où si les supports de $\hat{h}$ et $\hat{g}$ ne se recouvrent pas, $\widehat{C_{hg}}(\omega) = 0$. D'où le résultat.

Concrètement dans l'exemple de la figure 44, si l'on prend les coefficients d'une des imagettes notée $\bar{x}_1$ et son équivalent en $\bar{x}_j$ donc à une échelle 2^j où $j \neq 1$, alors le facteur de corrélation est nul. Alors que tout semblerait dire le contraire. Pourquoi a-t-on se phénomène? ***Il s'agit d'une interférence destructrice à cause de la différence de phase entre les deux objets***. Donc, ***si l'on ne fait que vouloir calculer la corrélation entre des mesures linéaires on aboutit à un échec***, car on obtient aucune information à travers les coefficients de corrélation.

Remarquons qui plus est que si X est un processus gaussien, $A = X * h$ étant une combinaison linéaire de v.a gaussiennes, c'est aussi une v.a gaussienne. Donc si $B = X * g$, on sait que A et B sont deux v.a gaussiennes de matrice de corrélation nulle pour g et h

103. NDJE. En écrivant le membre de gauche, et en se rappelant que les moyennes sont nulles, on obtient une double somme sur (v, v') par example dans la laquelle l'espérance entre selon $\mathbb{E}(X(v-u)X^*(v'-u') = C(v-v'+u'-u)$ dans le cas complexe pondérée de $h(v)$ et $g^*(v')$. En exprimant ce que vaut $(a*b*c)(x)$ d'une manière générale, on fait alors apparaitre la double convolution estimée en $u'-u$. D'où le résultat quand on utilise la définition de $\tilde{g}$.

ayant des supports disjoints, donc A et B sont ***deux v.a indépendantes*** et tous les moments d'ordre supérieurs sont également nuls. ***Donc, dans le cas gaussien la décorrélation signifie l'indépendance entre les v.a, d'où l'absence de structure***. Ce qui explique les résultats sur la générations de signaux (audio, ou turbulent) par des modèles gaussiens (Figs. 42, 43).

9.9 Réseau de scattering

Pour aller plus loin, et trouver un moyen de capturer la structure tout en utilisant des corrélations, il nous faut gérer les phases. ***On peut utiliser une non-linéarité*** de type $|x|$ ou bien [104] $ReLU(x)$.

Notons que si $\psi(u)$ a un spectre de Fourier qui est concentré autour de $\omega = \omega_0$, alors ψ_λ a son spectre concentré autour de $\omega = \lambda\omega_0$, et il en est de même de $X * \psi_\lambda$. Alors, si le on multiplie [105] par $e^{-i(\lambda\omega_0)u}$ le spectre se décale pour être concentré autour de $\omega = 0$. On peut procéder à ce décalage pour $X * \psi_{\lambda'}$ tout en laissant $X * \psi_\lambda$ si $\lambda' < \lambda$ afin que les deux spectres se recouvrent partiellement. Ce n'est pourtant pas suffisant pour annuler les interférences destructrices. Mais on peut maintenant utiliser la non-linéarité. Considérant $|X * \psi_{\lambda'}|$ et $X * \psi_{\lambda'}$, non seulement les spectres se recouvrent mais la corrélation est non nulle. Ainsi, on peut considérer les covariances suivantes, avec les échelles $\lambda = 2^j$ et $\lambda' = 2^{j'}$ telles que $j \neq j'$

$$\begin{cases} Cov((X * \psi_j)(u), |(X * \psi_{j'})(u - \tau)|) & = C^{(1)}(j, j', \tau) \\ Cov(|(X * \psi_j)(u)|, |(X * \psi_{j'})(u - \tau)|) & = C^{(2)}(j, j', \tau) \end{cases} \tag{323}$$

Dans le cas gaussien, on n'échappe pas au fait que $C^{(1)} = 0$ et $C^{(2)} = 0$ à cause de l'indépendance des variables mentionnée à la section précédente. Par contre, dans le cas non-gaussien on a *a priori* $C^{(1)} \neq 0$ et $C^{(2)} \neq 0$.

Remarquons que si on dispose d'un signal de N pixels (ou échantillons discrets), on a N valeurs de τ possibles, tandis que le nombre d'échelles [106] λ (ou λ') est de l'ordre de $\log_2 N$. Donc, ***le nombre de moments est de l'ordre de*** $N \log_2^2 N$. Par exemple si

104. Le ReLU annulant la partie négative des signaux fait effectivement l'affaire.
105. Voir par exemple Table 1 Sec. 3.4 Cours 2012.
106. NDJE. Voir par exemple Cours 2018 Sec. 6.6.0.3 les valeurs de j_{max}.

$N = 10^6$, cela fait 4 10^8 coefficients, donc beaucoup trop pour calculer autant de moments. Rappelons *primo* que nous disposons que d'1 seule image, et *secundo* nous voulons des estimateurs consistants qui demandent de faire des moyennes pondérées sur des lots de valeurs. Bref, ***il faut trouver un moyen de réduire le nombre de moments***.

Nous avons fait remarquer qu'étant donné que $X(u)$ est stationnaire, $|X * \phi_j|(u)$ sont des processus stationnaires. En pratique, ces coefficients d'ondelettes sont parcimonieux au sens où la quasi-totalité sont nuls exceptés ceux correspondants aux transitoires de $X(u)$. Donc, en prenant leurs modules, on a un nouveau signal que l'on peut analyser par ondelettes d'échelle j_2 en calculant les coefficients

$$(|X * \psi_j)| * \psi_{j_2})(u) \tag{324}$$

Il suffit alors de considérer les covariances suivantes

$$Cov((|X * \psi_j)| * \psi_{j_2})(u), (|X * \psi_{j'})| * \psi_{j_2})(u)) = C^{(3)}(j, j', j_2) \tag{325}$$

c'est-à-dire qu'il n'est pas nécessaire de faire intervenir le paramètres τ ce qui réduit drastiquement la combinatoire. C'est coefficients sont *a priori* non nuls uniquement dans la cas non gaussien.

Ces itérations (cascades) de convolutions par ondelettes forment ce que S. Mallat appelle des ***réseaux de scattering*** [107] (voir aussi Cours 2020 Sec. 9.5). D'un point de vue architectural de réseau, on peut se reporter [108] à la figure 45:

1. Pour l'image du haut. En partant d'une image x, on commence par appliquer une cascade du même type de celle présentée sur la figure 44, ce qui donne les imagettes vertes (filtrage passe-bas par ϕ) et les bleues (filtrage par les ϕ_j) reliées par les flèches vertes et noires. Ensuite, on applique à tous les coefficients des imagettes la non linéarité $\rho(x) = |x|$ dans le cas présent mais aurait pu être un ReLU. L'étape suivante est de considérer à présent chaque imagette bleue, une cascade identique à la précédente avec de nouveau un filtre passe-bas et une collection de filtres passe bandes à une nouvelle échelle j_2 (flèches rouges). Le notebook *scattering2D.ipynb* permet d'obtenir les coefficients de type $|x * \psi_\lambda| * \phi$ et $||x * \psi_\lambda| * \psi_{\lambda'}| * \phi$ en 2D.

107. NDJE. J. Bruna et S. Mallat *"Scattering Convolution Networks"* (2012) https://arxiv.org/pdf/1203.1513.pdf

108. NDJE. Je prends 2 images projetées par S. Mallat du même réseau afin d'éclairer la démarche.

L'application du filtre basse fréquence permet d'avoir des invariants par translation (Cours 2020 Sec. 9.4).

2. Pour l'image du bas. A chaque étape il faut renormaliser les coefficients (équivalent d'une BatchNorm) pour bien conditionner le problème. Une fois que l'on a les différents coefficients d'ondelettes, on peut calculer les corrélations à une échelle donnée (trait rouge), ce qui en langage des CNN correspondrait aux ***corrélations entre les canaux***. Ce sont ces moments que l'on va garder et ils sont au nombre de l'ordre de $K = \log_2^3 N$ soit environ 8,000 pour $N = 10^6$.

Ce type de réseaux de Scattering est une sorte de réseaux convolutionnels où les poids sont des filtres connus (cf. les ondelettes) à l'avance. Il n'y a pas d'apprentissage ici. On impose les ondelettes car on a de l'***information a priori***. On vaut capturer des ***transitoires*** donc on veut des ***filtres localisés en espace***; de plus les signaux étant ***stationnaires***, on veut des estimateurs de ***moments consistants***, donc il nous faut moyenner sur des plages de fréquences adaptées. Donc ***les filtres doivent également répondre à cette double exigence***. Inévitablement, on est amené à considérer les ***bases d'ondelettes.*** Ensuite, afin d'obtenir les interactions entre échelles, il nous faut s'affranchir des phases, d'où la nécessité des ***non-linéarité*** de type module ou ReLU. Donc, l'architecture du réseau est *in fine* naturelle.

Qu'est-ce que cela donne en pratique sur les images de type de celle du flux turbulent de la figure 43. Avec le réseau de scattering, on extrait les moments [109] $C^{(3)}(j, j', j_2) = \mathbb{E}(\phi_k(X))$ et on peut alors faire un modèle de $p(x)$ maximisant l'entropie ce qui donne

$$\tilde{p}(x) = Z^{-1} \exp\left(- \sum_{k=1}^{K} \beta_k \phi_k(x) \right) \tag{326}$$

On obtient les multiplicateurs de Lagrange β_k associés. Ensuite, on peut générer de nouvelles réalisations $\bar{x}$.

Les résultats sont montrés sur les figures 46 et 47. ***On arrive à capturer les structures contrairement au modèle gaussien***. Et quand les images originales sont de haute résolution, on peut tenir compte de plus de moments. Le modèle est encore plus fidèle et permet de générer des réalisations dont les statistiques des moments sont fidèles aux

109. NDJE. ici la covariance $Cov(a(u), b(u))$ signifie $\sum_u a(u)b(u)$ dans le cas où les moyennes sont nulles à cause de la renormalisation des coefficients.

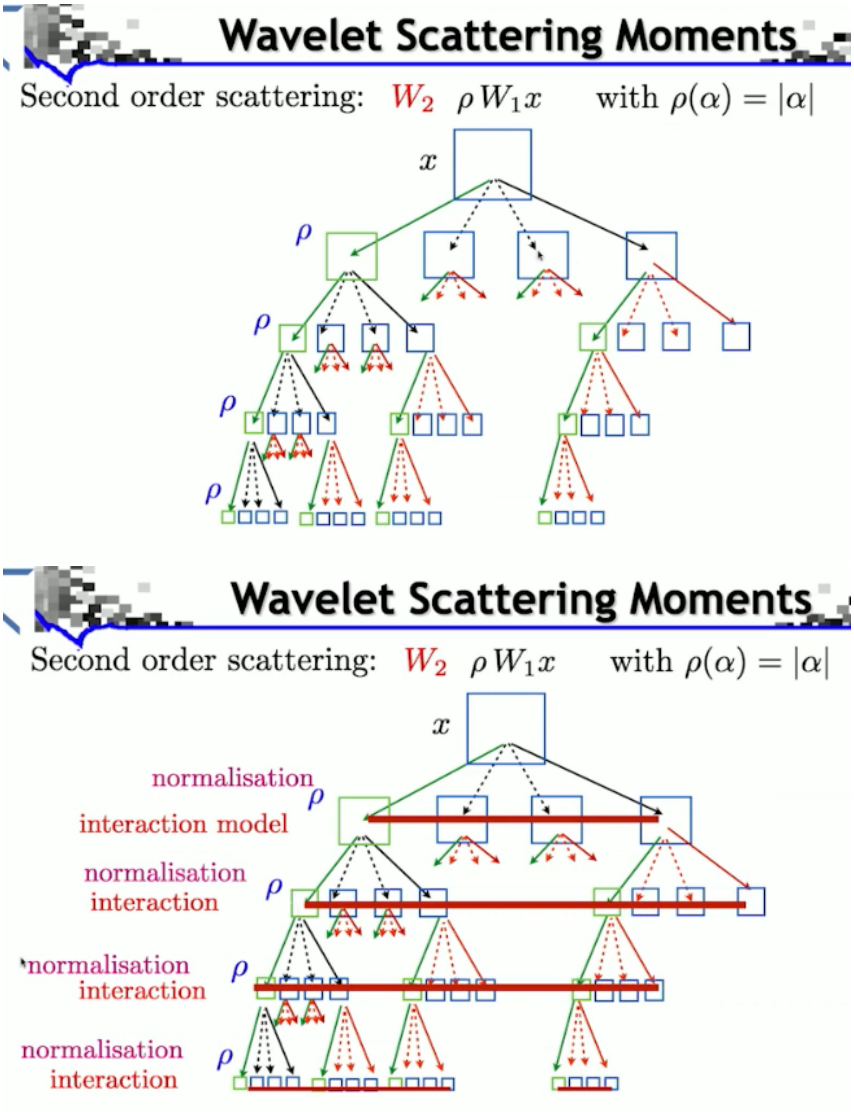


FIGURE 45 – Voir dans le corps du texte le descriptif des images.

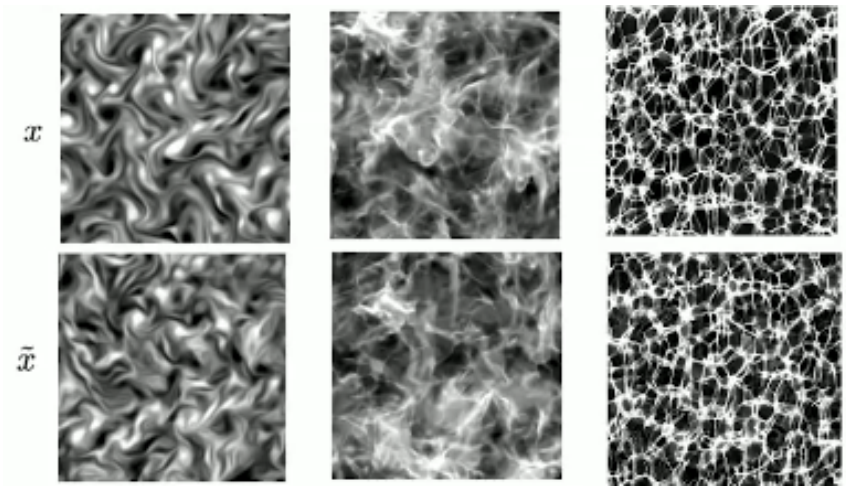


FIGURE 46 – Ligne du haut: images d'origines. Ligne du bas: nouvelles réalisations obtenues à partir de modèles d'entropie maximum élaborés à partir des coefficients $C^{(3)}(j, j', j_2)$ d'un réseau de scattering qui calcule les corrélations entre les canaux. Mais en se restreignant à $\log_2 N$ moments.

originaux jusqu'à l'ordre 4.

La description du modèle d'Ising est aussi non gaussienne à cause des interactions locales mais aussi parce qu'on impose des valeurs de spins ± 1. Là aussi la description par les réseaux de scattering qui fournissent les corrélations entre échelles permet de capturer la structuration multi-échelles du processus.

Dans le cas des sons, on obtient une nouvelle synthèse comme celle de l'image de droite de la figure 42. Ca marche aussi bien, à l'oreille on retrouve les structures des différents sons originaux.

10. Épilogue

Y'a-t'il une limite à la description d'un phénomène par les coefficients d'ordre 2 obtenus par les réseaux de scattering qui rappelons le n'ont aucun poids appris? Qu'en est-il des processus tels que les visages? la réponse courte est: ***tout ce que l'on a évoqué dans***

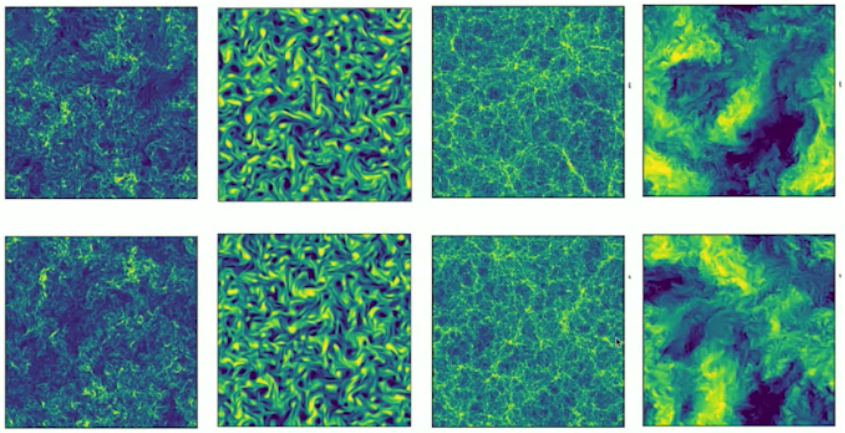


FIGURE 47 – Même type d'exercice (Fig. 46) de génération de nouvelles réalisations à partir des moments d'ordre 2 calculés par un réseau de scattering, mais en prenant plus de coefficients car les images originales sont de plus haute résolution. Dans ce cas on peut reproduire les statistiques jusqu'aux moments d'ordre 4.

ce cours s'effondre. La question est pourquoi? L'hypothèse d'ergodicité est fausse en particulier à cause de la non stationnarité (on recentre toutes les images). Mais mettons qu'1 image de visage est une réalisation du processus "visage", les propriétés de ce processus sont bien plus complexes que celles des processus physiques envisagés précédemment.

Faisons un petit bilan du cours. ***La notion centrale est celle d'entropie***. On peut la voir comme une mesure d'incertitude, mais nous savons qu'il s'agit d'un comptage des états discrets que le système peut prendre. Et nous avons pu définir pour des ***processus stationnaires un taux d'entropie*** qui nous a particulièrement servi dans la cas de chaine de Markov (Sec. 7.3). Mais dans le cas d'un système qui devient de plus en plus grand ($n \to \infty$) pour lequel on a n'a pas ces hypothèses d'ergodicité/stationnarité, il devient alors très compliqué de prévoir ce qu'il se passe. En particulier, l'entropie va-t-elle fluctuer ou bien converger?

On sait que si l'***entropie converge*** alors on peut disposer de beaucoup de résultats. Par exemple, on observe la répartition asymptotique dans les espaces typiques avec une probabilité quasi uniforme, et le nombre d'éléments de ces ensembles est donné par l'entropie. Donc l'entropie est la variable qui est au cœur de la description du système. On a vu aussi, que l'évolution temporelle d'un tel système, converge vers un maximum

d'entropie (2nd principe de la thermodynamique) correspondant à la mesure stationnaire (équilibre) dans le cas markovien. In fine, on peut résumer en disant que dans ce cas ***on a un contrôle sur les propriétés mathématiques du système***. Dans le cas gaussien, c'est encore plus simple car il y a une absence de structure; dans le cas non gaussien durant les 15 dernières années nous dit S. Mallat, on a pu se convaincre que l'on peut également maitriser le sujet (cf. les réseaux de scattering) en comprenant les interactions d'échelles.

Par contre, dans le cas des visages que faire? Comme le dit S. Mallat, on peut se demander après tout s'il y a un processus sous-jacent à la "production" d'une image de visage? Si tel n'est pas le cas, on ne peut appréhender les propriétés statistiques. A un certain moment, il faut pouvoir calculer des espérances, c'est-à-dire des moyennes empiriques dans le cas pratique, donc disposer de plusieurs échantillons issus du même processus... Mais si chaque image avait sa propre "histoire", ce ne serait pas envisageable. Or, ***nous disposons d'algorithmes*** (voir le séminaire de Valentin de Portoli 2023) ***qui fonctionnent d'une manière spectaculaire***: voir les architectures Generative Adversarial Network, Variational Auto-Encoder ou de diffusion... Donc, ***il y a vraisemblablement un modèle mathématique sous-jacent***. C'est le domaine actuel de la recherche.

Quand on regarde les résultats de ces modèles génératifs, on constate qu'il y a un ***phénomène de mémorisation***, à savoir qu'il y a des parties de l'image générée qui se retrouvent dans la base de données. Ce qui n'est pas du tout ce que l'observe quand on procède selon le schéma de modélisation par entropie maximum. Quand on procède ainsi, on favorise la diversité puisque l'on diffuse la probabilité sur l'ensemble des éléments d'un ensemble typique. Alors que dans les modèles GAN, VAE, etc, il semble que ce n'est pas la diversité qui soit maximisée, ce n'est pas non plus l'exemple trivial de tirer au hasard un élément de la base de données. Il y a de la variabilité sans maximiser l'entropie, car on retrouve ces structures présentes dans la base données. Clairement un modèle guidé par le principe d'entropie maximum aurait une probabilité nulle de générer à nouveau ne serait-ce qu'une petite portion de l'image à l'identique de celle présente dans la base de données. ***Il faut donc renoncer au principe d'entropie maximum.*** Notons que si l'on force ces modèles génératifs à augmenter l'entropie, la qualité des images se dégrade, les fines textures disparaissent (ex. les cheveux).

Selon S. Mallat, il faut donc essayer de bâtir un pont entre une modélisation simple de mémorisation de visages, et la modélisation de la probabilité par le principe d'entro-

pie maximum qui impose des propriétés fortes d'ergodicité du processus. C'est un des thèmes abordés par Marc Mézard (*Physique statistique et inférence: le défi des données structurées*) lors du dernier séminaire de cette année.

Pour finir, S. Mallat nous relate qu'environ 99% des activités des chercheurs, doctorants sont centrées sur la mise au point d'architectures afin de rivaliser d'audace pour augmenter les performance., Il y a pourtant un autre aspect tout aussi passionnant, c'est celui d'essayer de comprendre ce qui fait le succès de telle ou telle architecture.

NDJE. (extrait traduit et légèrement adapté de la thèse de John Zarka). "Trouver un bon équilibre entre la théorie et la pratique est toujours un exercice subtil. Certains cadres théoriques peuvent bien expliquer les propriétés de base d'architectures profondes, mais sont limités dans leurs applications pratiques, ou reposer sur des hypothèses trop restrictives. D'un autre côté, les architectures profondes récentes qui obtiennent l'état de l'art sur des problèmes à grande échelle ont atteint un tel niveau de complexité technique qu'il semble difficile d'ouvrir ces boîtes noires et d'exposer les fondements mathématiques qui conduisent à leurs performances impressionnantes."

Donc le message de S. Mallat est *"on peut essayer de faire moins bien tout en tentant de comprendre mieux."*.