Notes et commentaires au sujet des conférences de S. Mallat du Collège de France (2022)

Jean-Eric Campagne

Notes et commentaires au sujet des conférences de S. Mallat du Collège de France (2022)

Modèles multi-échelles et réseaux de neurones convolutifs

J.E Campagne *

Janv. 2022; rév. 19 février 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 également que sur le site associé à ses cours S. Mallat*[1] *donne en libre accès des chapitres de son livre "A Wavelet Tour of Signal Processing", 3ème edition.* ainsi que d'autres matériels.

Cette année 2022 c'est la cinquième du cycle de la chaire de la Science des Données de S. Mallat, le thème en est: la **Théorie de l'Information**.

J'ai mis quelques notebooks sur github[2] *pour illustrer ce cours. Cette initiative est minimaliste et donc vous êtes invités à me faire un retour et des propositions. J'ai utilisé JAX comme librairie d'auto-differentiation, car elle permet de coder directement à-la-Numpy ce qui facilite l'apprentissage.*

En ce début de Mars 2022, la vague du Cocid-19 omicron qui a sévi en début d'année tend à disparaître, malheureusement je ne peux passer sous silence ce qu'il faut bien appeler la Guerre en Ukraine déclenchée le 24 février par le Président V. Poutine et qui va changer le contexte dans lequel nous vivons.

2. Séance du 19 Janv.

2.1 Introduction

Faisons le point sur quelques faits marquants dans le domaine de la science des données en 2021. On peut citer par exemple la reconnaissance de la performance de très grands systèmes tels que GPT-3 développé par Open AI [3] et mis en service à la mi 2020. Ce

1. https://www.di.ens.fr/~mallat/CoursCollege.html
2. https://github.com/jecampagne/cours_mallat_cdf/cours2022
3. https://openai.com/blog/openai-api/, Tom B. Brown et al. *Language Models are Few-Shot Learners*, (Juil. 2020) arXiv:2005.14165v4 https://arxiv.org/abs/2005.14165

système a la bagatelle de 175 milliards de paramètres et constitue à date le plus gros. C'est un modèle de langage formel dont la base d'apprentissage se nourrit de base de données tirées du Web comme Common Crawl, WebText2 [4], de Google Books et Wikipedia. Il est donc entrainé sur des centaines de milliards de mots. La tendance de fond à laquelle on assiste depuis le début des réseaux de neurones, c'est que plus les modèles ont de paramètres, plus les performances deviennent spectaculaires. Qui plus est GPT-3 n'est pas confiné dans une tâche particulière, d'un corpus particulier, il devient d'une certaine manière généraliste car il est capable de générer tout type de textes (ex. traduction dans n'importe quelle langue à partir d'un seul exemple, arithmétique, n'importe quel langage informatique, d'écrire des textes à partir d'un exemple) mais aussi de dialoguer, etc. Des humains ont de plus en plus de mal a détecter l'origine artificielle ou humaine des articles même de plus de 200 mots. Malheureusement, le revers de la médaille est la porte ouverte à la désinformation, aux messages frauduleux d'une manière générale totalement générés automatiquement.

Maintenant, le domaine reste très expérimental, et ces performances sont mal comprises bien que la "découverte" de *la double descente du risque* par Belkin et al. [5] dont S. Mallat a parlé dans son cours de 2020 [6] génère beaucoup de pistes d'étude dans le domaine de la *sur*-paramétrisation. Il y a une profusion de publications (ex 15,000 papiers à la dernière conférence NISP), une accélération de la recherche. Et pourtant, il y a besoin de revenir à des bases pour avoir une perspective globale et si l'on a tendance à penser ou constater que des articles sont obsolètes au bout de quelques mois, il y en a qui traversent les siècles. Par example c'est autour des années 1920 que **Ronald A. Fisher** (1890–1962) pose les bases de la *Statistique*, et finalement on est en plein dans le programme qu'il établit le 1er Janvier 1922 "*On the mathematical foundations of theoretical statistics*" [7]. Il en est de même de l'article de 1948 de **Claude Shannon** (1916-2001) *"A Mathematical Theory of Communication"* [8].

4. https://commoncrawl.org/, https://www.eleuther.ai/projects/open-web-text2/

5. Mikhail Belkina, Daniel Hsub, Siyuan Maa, and Soumik Mandala, "Reconciling modern machine learning practice and the bias-variance trade-off", arXiv:1812.11118v2

6. note J.E.C, Notes et commentaires au sujet des conférences de S. Mallat du Collège de France (2020), Modèles multi-échelles et réseaux de neurones convolutifs, Fev 2020; rév. 17 septembre 2020. https://www.di.ens.fr/ mallat/CoursCollege.html

7. https://doi.org/10.1098/rsta.1922.0009 disponible sur le site du cours https://www.di.ens.fr/~mallat/CoursCollege.html

8. C. E. Shannon, The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October.

Nous allons donc étudier durant ce cours la notion d'**Information**. Cependant, avant cela on va se demander qu'est-ce que l'on entend par construire un modèle (ex. réseaux de neurones), et quel type de modèle(s) peut-on choisir quand on veut faire de l'analyse en très grande dimension?

2.2 Modèle déterministe vs stochastique

Sous ce choix se joue en filigrane une certaine vision du monde [9], avec d'un côté une perspective cartésienne plutôt française (continentale), et de l'autre une vision bayésienne plutôt anglaise [10]. Disons que ces deux visions des probabilités ont leurs biais respectifs, si on veut faire court. Si culturellement, on peut vouloir pencher d'un côté ou d'un autre et se dire que les deux visions sont équivalentes, pour ce qui concerne des problèmes à grande dimension on n'a pas le choix en quelque sorte.

Prenons le problème de *classification supervisée*. L'objectif est d'estimer une réponse y à partir de données $x \in \mathbb{R}^d$ ($d \gg 1$), et pour ce faire on dispose d'un lot d'entrainement $\{x_i, y_i\}_{i<n}$. La question qui se pose est alors: est-ce plus difficile de résoudre ce problème quand d augmente?

Prenons le point de vue *déterministe*: la réponse est **Oui** à cause de *la malédiction de la dimensionalité* qui était le sujet notamment du Cours de 2018. Si l'on considère une fonction inconnue $y = f(x)$ en 1D, et si l'on dispose de suffisamment de points (d'échantillonnage) et que la fonction f est suffisamment régulière, alors on va pouvoir l'interpoler avec une bonne approximation (Fig 1). Maintenant, quand on se place en grande dimension $x \in \Omega$ (ex. $\Omega = [0,1]^d$) si on veut des données suffisamment denses, par exemple avec une distance ε entre points adjacents, il faut N points pour couvrir l'espace Ω. On aboutit alors à la relation d'échelle suivante

$$N \sim \varepsilon^{-d} \tag{1}$$

https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf.

9. Voir par exemple Cours 2019 Sec. 2.3.2

10. Cependant, nous verrons l'apport de Pierre-Simon de Laplace (1749-1827) qui redécouvre la loi des probabilités inverses de Bayes donnera naissance en 1812 à une *Théorie des Probabilités* dont les éléments forment également la base des travaux actuels.

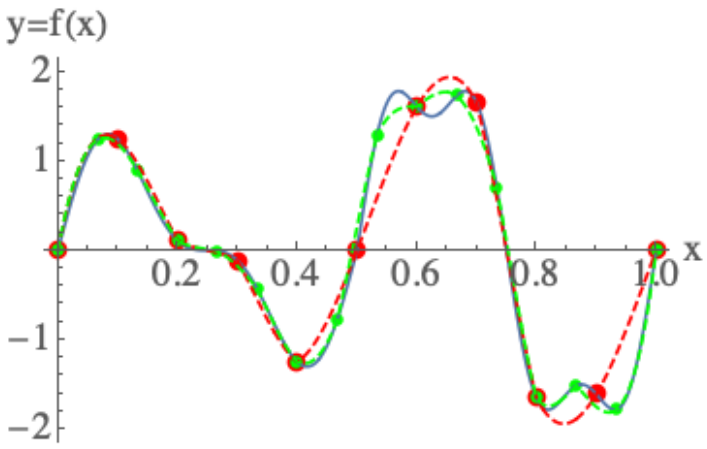

FIGURE 1 – Interpolations de deux jeux de données (x_i, y_i) (ici non bruitées; points rouges et verts) d'une fonction f sous-jacente inconnue (courbe bleue). Plus la population d'échantillons est dense meilleure est l'interpolation.

Or, si l'on a une fonction régulière, par exemple lipschienne, alors quand on se trouve au voisinage d'un point d'entrainement

$$\|f(x) - f(x_i)\| \leq C\|x - x_i\| \tag{2}$$

et la densité de point permet de borner le membre de droite, et donc d'estimer l'erreur de généralisation (membre de gauche). Ainsi, le nombre de données nécessaires N explose exponentiellement avec la dimension d pour maintenir la précision de l'ajustement. C'est ce phénomène d'explosion qui motive la réponse. Ceci dit on pourrait toujours argumenter sur la régularité des fonctions sous-jacentes des données en grande dimension, pour se dire que le problème peut être attaqué malgré tout: par exemple en exprimant des invariants/symétries du problème (ex. la thématique du Cours de 2020) pour faire de la réduction de dimensionalité. Cependant, dans le raisonnement ci-dessus, on s'aperçoit qu'il n'y a pas de modèle sur les données. C'est là que l'approche stochastique va d'une certaine manière tenter d'aller plus loin dans l'analyse.

Prenons le point de vue *bayésien*: la réponse est **Non**! En effet quand d augmente, intuitivement on constate qu'une image a une meilleure résolution (idem pour un extrait sonore), et donc il parait naturel qu'il soit plus facile de reconnaitre *a priori* un objet

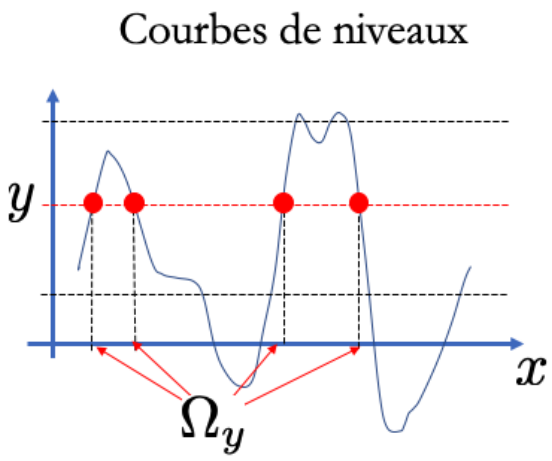

FIGURE 2 – Exemple d'une ligne de niveau $f(x) = y$.

mieux résolu dans l'image. Ainsi, la réponse négative est naturelle, et on a donc un malaise. Envisageons les **lignes de niveau** [11] de la fonction f (Fig. 2):

$$\Omega_y = \{y \ / \ f(x) = y\} \tag{3}$$

Ce qui nous intéresse c'est la ***géométrie*** de ces lignes (en dimension quelconque il s'agit de surfaces). Où les points se concentrent-ils dans l'espace? Or, il y a bien concentration (Fig. 3) et l'espace "réellement" occupés par exemple par les images de chiens, chats, voitures, etc est tout petit par rapport à l'espace total possible des images quelconques de même dimensionalité. Le phénomène sous-jacent est ***la loi des grands nombres***. Finalement, avoir la vision des lignes de niveaux, c'est d'une certaine manière opter pour une vision similaire à celle de l'intégrale de Lebesgue qui utilise *la mesure de ces ensembles*. Et qui dit *mesure* dit *probabilité*. Schématiquement, via une mesure on dispose de la densité de probabilité de x sachant y, soit $p(x|y)$:

$$\Omega_y \xrightarrow{\text{mesure}} p(x|y) \tag{4}$$

11. Argumentaire élaboré dans le Cours de 2021 à propos du théorème de A. Barron de 1993 (Sec. 5.2.3).

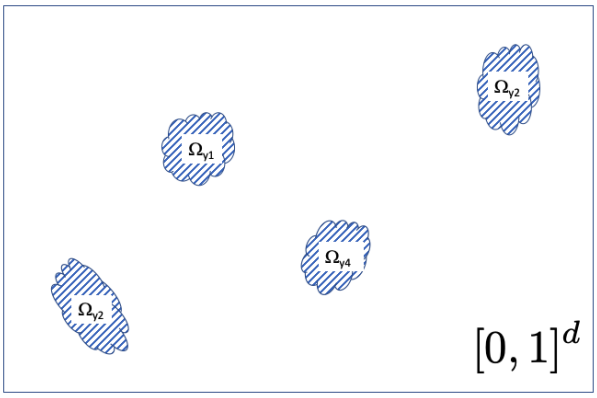

FIGURE 3 – Concentration des classes d'objets.

Or, via le théorème de Bayes (Thomas Bayes 1701-61), nous avons

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{5}$$

où $p(y)$ et $p(x)$ sont les probabilités *a priori*[12], et $p(x|y)$ est appelée la vraisemblance (*likelihood*) que x soit vraie sachant y. Le classificateur bayésien définit le meilleur y comme celui qui ***maximise*** $p(y|x)$:

$$\hat{y} = \underset{y}{\text{argmax}}\ p(y|x) \tag{6}$$

Pour réaliser ce schéma, l'approche bayésienne se pose (ou doit se poser) la question où les données se concentrent-elles? et alors on réalise que l'on n'a pas besoin de trouver une solution y pour n'importe quel $x \in \Omega$ mais bien uniquement pour les x éléments des Ω_y.

Donc, il faut aborder ***les phénomènes de concentration*** de la mesure, et modéliser des probabilités $p(x|y)$ (problème non-supervisé) ou $p(y|x)$ (problème supervisé). Typiquement, on va étudier des familles de probabilités telle l'exponentielle, et par exemple la modélisation selon

$$p(x|y) = Z_y^{-1} e^{\Theta_y . \Phi(x)} \tag{7}$$

revient à se poser la question de modéliser $\Phi(x)$, c'est-à-dire **la représentation de** x la

12. On parle de *prior* pour $p(y)$ et de *marginal likelihood* pour $p(x)$ car on peut écrire que $p(x) = \int p(x|y)p(y)dy = \int p(x,y)dy$.

plus appropriée qui linéarise $\log p(x|y)$.

On se rend compte alors que les domaines qui étudient ce type de probabilités, sont la ***Physique Statistique*** et la ***Théorie de l'Information*** que nous allons voir durant ce cours de 2022.

2.3 Le point de vue de Fisher

La première question que R. Fisher se pose dans l'article de 1922 est de savoir comment définir l'information des données sur l'estimation d'un paramètre θ? C'est un problème *d'**Inférence***. Il fait toute une réflexion sur qu'est-ce que l'on essaye de faire dans ce domaine des mathématiques statistiques et d'analyse des données. Selon lui on essaye de pratiquer une forme de ***compression des données***, c'est-à-dire représenter les données avec le moins de paramètres possibles, et fournir une représentation de l'information importante avec les données dont on dispose. Il développe alors

- la notion de ***consistance*** d'estimateur [13]; est-ce que l'estimateur converge quand on a une infinité de données, est-il biaisé ou pas?
- la notion ***d'inférence*** par maximum de vraisemblance
- la notion ***d'information***
- la notion de ***statistique suffisante*** ou *exhaustive* qui rend compte du fait que la statistique (ensemble d'opérations appliquées à un jeu de données) contient toute l'information sur le ou les paramètres de la loi de probabilité sous-jacente.

Toutes ces notions forment la base du programme que constituent les mathématiques statistiques actuelles.

Pour fixer les idées, si l'on dispose d'un jeu de données $\chi = \{x_i\}_{i\leq n}$, le problème posé est de connaître la loi de probabilité qui sous-tend la naissance de ce jeu particulier. Ainsi, R. Fisher définit une ***famille de probabilités*** indicées par θ, $p_\theta(x)$, et le problème revient à estimer le "bon" θ. Par exemple en 1D, on peut penser à $\theta = (\mu, \sigma^2)$ tel que

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8}$$

13. NDJE: il se trouve qu'en anglais on utilise le vocable "the parameter consistency" qui peut plutôt être traduit par "la cohérence d'estimateur", mais j'utiliserai le vocable de S. Mallat.

Bien entendu, le problème que l'on a en tête avec la classification d'images par exemple se pose avec beaucoup plus de variables.

Soit l'estimateur $\hat{\theta}(\chi)$, on aimerait que selon un sens à trouver, on obtienne une convergence de l'estimateur quand $n = |\chi|$ tend vers l'infini, c'est-à-dire

$$\hat{\theta}(\chi) \xrightarrow[|\chi|\to\infty]{} \theta \tag{9}$$

alors on qualifie $\hat{\theta}(\chi)$ d'***estimateur consistant***. Fisher va alors trouver un moyen de construire des estimateurs consistants, c'est le ***maximum de vraisemblance***:

$$\hat{\theta}(\chi) = \underset{\theta}{\operatorname{argmax}}\ p_\theta(\chi) \tag{10}$$

En cela, on détermine *un modèle pour lequel les données observées sont les plus probables possibles*. Par la suite, on notera l'estimateur $\hat{\theta}$ sous-entendu les observations χ. Lorsque l'on dispose d'*observations identiquement distribuées et indépendantes* (*iid* par la suite), alors

$$p_\theta(\chi) = \prod_{i\leq n} p_\theta(x_i) \tag{11}$$

Il est tentant d'utiliser le logarithme, on définit alors le *log-likelihood* (on supprimera parfois le log pour ne parler que du *likelihood*)

$$\ell(\theta) := \log p_\theta(\chi) = \sum_{i\leq n} \log p_\theta(x_i) \tag{12}$$

et donc l'idéal est de trouver le $\theta(\chi)$ qui maximise l'espérance du likelihood.

$$\hat{\theta}(\chi) = \underset{\theta}{\operatorname{argmax}}\ \mathbb{E}_\chi[\ell(\theta, \chi)] \tag{13}$$

Dans ce contexte, la ***notion d'indépendance*** (des observations) est ***la forme de régularité*** qui va permettre *in fine* de vaincre la malédiction de la dimension.

Concernant l'***Information de Fisher***, c'est l'idée de calculer l'incertitude sur le paramètre (et de la propager à l'erreur d'estimation de généralisation). Comme $\ell(\hat{\theta})$ est maximale, alors

$$\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0 \tag{14}$$

et l'on peut regarder si le maximum est plus ou moins "étroit" en se servant de la dérivée d'ordre supérieur par exemple (notion de *courbure*). Une autre façon d'aborder le problème, si l'on a affaire à un ***estimateur non-biaisé***, c'est-à-dire[14] $\mathbb{E}(\hat{\theta}) = \theta$, c'est de regarder la variance (sachant que $\mathbb{E}[\partial\ell(\hat{\theta})/\partial\theta] = 0$)

$$\boxed{I(\theta) = \mathbb{E}\left[\left(\frac{\partial\ell(\hat{\theta})}{\partial\theta}\right)^2\right]} \tag{15}$$

que Fisher appelle *information*[15]. Le résultat de Cramér-Rao[16] donne une borne supérieure de l'erreur d'estimation

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \geq \frac{1}{I(\theta)} \tag{16}$$

qui donne un sens au fait que ***plus on a d'information (utile) sur le paramètre θ meilleure est son estimation***.

Il faut reconnaitre que tout ce schéma développé par R. Fisher, est bien ce qu'on essaye de faire quand on effectue une descente de gradient stochastique pour entrainer un réseau de neurones.

2.4 Le cas des réseaux de neurones

Pourquoi le schéma de Fisher est à l'œuvre dans le cas d'optimisation des réseaux de neurones? Le problème n'est pas tellement de développer le formalisme décrit précédemment, le grand problème est de *spécifier la famille de probabilités $p_\theta(x)$*. Les réseaux de neurones peuvent être vus comme une façon de spécifier la dite famille.

Par exemple, sur la figure 4 on a schématisé différentes étapes typiques d'un réseau de neurones. On a une cascade de filtres linéaires (ex. convolution), de non-linéarités (ex. ReLU), pour finir par une opération linéaire qui donne un vecteur $z_y(x)$ duquel, par une

14. L'espérance est à prendre dans le sens où l'on se donne une loi de génération de jeu d'observations χ, ce qui donne un caractère aléatoire à $\hat{\theta}(\chi)$ dont on peut calculer l'espérance, la variance, etc.

15. NDJE: il s'agit bien de la variance impliquant la dérivée première de $\ell(\theta)$, mais si la fonction est deux fois dérivable alors on a l'espérance de $-\partial^2\ell/\partial\theta^2$ avec le changement de signe qui convient.

16. Harald Cramér (1893-1985) et Calyampudi Radhakrishna Rao (1920-).

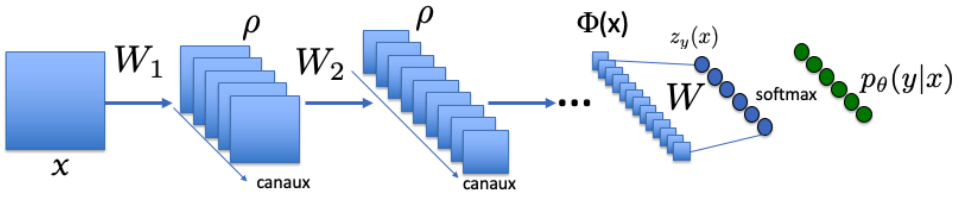

FIGURE 4 – Schématisation d'un réseau de neurones multi-couches (classification).

opération "`softmax`" [17], on obtient la loi de probabilité $p_\theta(y|x)$ selon

$$p_\theta(y|x) = \frac{e^{z_y(x;\theta)}}{\sum_{y'} e^{z_{y'}(x;\theta)}} \tag{17}$$

avec y' courant sur l'ensemble des classes à séparer (ex. les digits de 0 à 9). Dans ce contexte, les paramètres θ sont tous les coefficients des filtres convolutionnels et de la linéarité finale. L'estimateur de la sortie du réseau, ici noté $\hat{y}$, réalise le maximum de probabilité

$$\hat{y} = \underset{y}{\operatorname{argmax}}\ p_\theta(y|x) \tag{18}$$

Pour optimiser la classification, on calcule simultanément $\hat{\theta}$ comme le maximum de la vraisemblance qui nous le verrons est équivalent à minimiser la "distance" Kullback-Leibler [18], c'est-à-dire une entropie conditionnelle. Si l'on note $\mathcal{D} = \{x_i, y_i\}_{i \leq n}$

$$\hat{\theta} = \underset{y}{\operatorname{argmax}} \underset{\{x,y\}\sim\mathcal{D}}{\mathbb{E}}[\log p_\theta(y|x)] \tag{19}$$

On a donc une fonction de coût que l'on minimise $(-\ell(\theta))$, et pour cela on utilise une descente de gradient [19].

Le point remarquable est le constat que les familles de probabilités auxquelles les réseaux de neurones permettent d'accéder sont assez génériques pour permettre de résoudre des classes de problèmes très larges comme de l'imagerie, du traitement de texte, de l'audio, de la physique/chimie etc. Un point qui nous occupe alors est de comprendre la

17. Voir ex. Cours 2020 Sec. 3.4.
18. Voir Cours de 2019 Sec. 7.2.3 par exemple.
19. Voir Cours 2018 Sec. 10 et 2019 Sec. 8 par exemple.

nature de ces familles de probabilités et en quoi ces modèles neuronaux sont-ils complexes.

2.5 Autre information: celle de Shannon

L'information de Fisher développée jusqu'à présent, est basée sur l'*a priori* que l'on a un modèle paramétré, et on essaye d'inférer les meilleurs paramètres possibles en se donnant un critère. Un autre type totalement différent d'information a été mis en œuvre par **Claude Shannon** (1916-2001) dans les années 1940. C'est ***une information qui ne dépend plus du modèle***. L'idée est de se demander quelle est l'*information intrinsèque* contenue dans les données? En sous-jacent ce sont des problématiques de communications entre émetteur-récepteur qui ont donné le terreau à ces développements, car il faut préserver le maximum d'information dans ces échanges. L'article fondateur date de 1948 et son titre est similaire à celui de R. Fisher: *A Mathematical Theory of Communication*[20]. Comme son ainé, Shannon écrit[21] là un article qui éclaire tout un domaine toujours d'actualité: le pourquoi, les nouveaux outils et les théorèmes de base.

Le cadre est le même que précédemment: l'on a une série d'*observations indépendantes* $\chi = (x_i)_{i\leq n}$, produites par une même loi de probabilité[22] donc se sont des données *iid*. Si l'on note $p(x_i)$ la probabilité de l'observation x_i, alors

$$p(\chi) = \prod_{i=1}^{n} p(x_i) \Longrightarrow \frac{1}{n}\log p(\chi) = \frac{1}{n}\sum_{i} \log p(x_i) \tag{20}$$

Le membre de droite représente une moyenne des log-probabilités des x_i. Or, les variables sont indépendantes, donc la loi des grands nombres nous donne (si tout se passe bien) une convergence quand n tend vers l'infini

$$\boxed{\lim_{n\to\infty} -\frac{1}{n}\log p(\chi) = \mathbb{E}[-\log p(x)] := \mathbb{H}[p]} \tag{21}$$

20. 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.

21. également à 32ans comme Fisher.

22. S. Mallat signale qu'il utilise la notion de probabilité selon la mesure de Lebesgue, mais on peut également avoir en tête la notion de mesure directement.

où $\mathbb{H}$ est l'***Entropie de Shannon***[23]. La propriété qui se dégage ici, indépendamment qu'il y ait un modèle, est que ***la probabilité d'un ensemble d'observations a tendance à converger***.

Explicitons cette propriété: le fait que cela converge en probabilité, cela veut dire que $\forall \varepsilon > 0$, on a

$$\mathbb{P}\left(\left|-\frac{1}{n}\log p(\chi) - \mathbb{H}[p]\right| \leq \varepsilon\right) \xrightarrow{n\to\infty} 1 \tag{22}$$

Donc les observations x en réalité, n'occupent pas vraiment tout l'espace Ω mais se concentrent dans un espace, dit ***espace typique*** T^ε définit par le fait que (Fig. 5)

$$\boxed{T^\varepsilon = \left\{\{x\} \in \Omega \ / \ \left|-\frac{1}{n}\log p(\{x\}) - \mathbb{H}[p]\right| \leq \varepsilon\right\}} \tag{23}$$

Cet ensemble est potentiellement beaucoup plus petit que l'ensemble Ω, sa taille est donnée par l'entropie $\mathbb{H}$. En quelque sorte, l'entropie définira *le nombre de bits minimum* qui vont permettre de coder les observations. On se retrouve avec une *notion d'information*, dont l'origine est plutôt une *notion d'incertitude*, reliée à la taille de l'ensemble typique. Le point remarquable est que *la densité de probabilité à l'intérieur de l'ensemble typique est uniforme*. On se retrouve finalement dans un problème de géométrie, car caractériser les observations revient à caractériser la géométrie de l'ensemble typique.

L'impact de ces notions est formidable car il en découle toute l'industrie des télécom. (codage, capacité d'un canal), et à nouveau cela se retrouve en Physique Statistique à travers la notion d'entropie et d'ensembles typiques. En mathématiques, quand on veut regarder la probabilité d'événements rares, on se sert à nouveaux de l'entropie[24]. Les

23. NDJE: Un peu d'histoire sur la notion d'entropie même si l'exercice ne peut être exhaustif. Depuis les travaux de Rudolf Clausius (1822-88) à qui l'on doit le premier concept d'entropie (1865), puis ceux de J. Clerk Maxwell (1831-79) qui développa la théorie de la distribution des vitesses dans les gaz, généralisée en 1896 par Ludwig Boltzmann (1844-1906) qui interprète l'entropie selon la fameuse formule "S = k log W " gravée sur sa tombe, la Mécanique Statistique s'appuie sur les travaux de Josiah Willard Gibbs (1839-1903). Il écrit en 1901 un ouvrage "*Elementary Principles in Statistical Mechanics developped with especial reference to the Rational Foundation of Thermodynamics*" (Yale Univ. publié en Mars 1902), établissant un pont solide entre la Mécanique Statistique et la Thermodynamique, et généralise l'interprétation statistique de l'entropie d'un système que reprend Claude Shannon en 1948.

24. NDJE: S. Mallat fait référence à la *Théorie des Grandes Déviations* développée dans les années 60s dans la lignée de C. Shannon, par entre autres des auteurs comme Harald Cramér (1893-1985), S. R. Srinivasa Varadhan, Jürgen Gärtner, Richard S. Ellis, Ivan Nikolaevich Sanov (1919-1968), et Edwin Thompson Jaynes (1922-98).

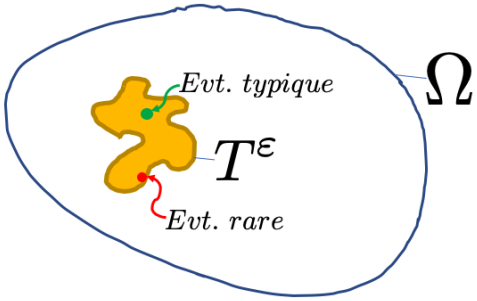

FIGURE 5 – Schématisation d'un ensemble typique T^{ε} dont la probabilité d'appartenance tend vers 1 quand le nombre d'observations tend vers l'infini. La taille de l'ensemble typique est essentiellement proportionnelle à l'entropie de Shannon $\mathbb{H}$. En vert une observation typique, en rouge une observation rare qui se trouve donc à la bordure de T^{ε}.

événements rares se retrouvent à la bordure des ensembles typiques.

On a donc là aussi une très belle théorie développée par Shannon et ses successeurs sauf qu'il faut être capable de caractériser les ensembles typiques. Si dans le cas de Fisher on a une paramétrisation explicite, dans le cas de Shannon il faut une géométrie, mais quelle est donc cette géométrie? Un cas qui a été étudié très en détail en premier lieu car il est plus facile, c'est celui des ***Processus Gaussiens***.

2.6 Le cas des Processus Gaussiens

D'un certain coté ces processus gaussiens sont les équivalents chez Shannon de la paramétrisation de la famille de gaussiennes chez Fisher. Ainsi, notons la probabilité jointe selon [25]

$$p_\theta(x) = Z^{-1} \exp\left\{-\frac{1}{2} x^T \Theta^{-1} x\right\} \tag{24}$$

où Θ est la matrice de covariance du processus gaussien (supposé de moyenne nulle):

$$\Theta = \mathbb{E}(xx^T) \tag{25}$$

25. nb. $x^T\Theta^{-1}x$ pourra être noté $\langle x, \Theta^{-1}x\rangle$.

(x un vecteur de dimension $d \times 1$, donc Θ est de dimension $d \times d$). Dans ce contexte, à quoi correspondent les ensembles typiques? Pour cela, il nous faut étudier la log-probabilité qui ici est très simple:

$$-\log p_\theta(x) = \log Z + \frac{1}{2} x^T \Theta^{-1} x \tag{26}$$

L'idée immédiate qui vient à l'esprit est de diagonaliser la matrice Θ et d'en obtenir les valeurs et vecteurs propres. Alors on peut écrire

$$\frac{1}{2} x^T \Theta^{-1} x = \sum_{k=1}^{d} \frac{x^2(k)}{2\sigma_k^2} \tag{27}$$

où les $(x(k))_{k \leq d}$ sont les d coordonnées de x dans une base qui diagonalise la matrice de covariance telle que dans cette base $\Theta = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. En définitive on effectue une analyse en composantes principales (PCA). ***Les ensembles typiques sont donc caractérisés par des ellipsoïdes dont les axes de symétries sont les axes principaux de la matrice de covariance***.

Cependant, même si le théorème central limite a plutôt tendance à donner crédit à l'utilité des Processus Gaussiens, il n'en reste pas moins vrai que les problèmes réels sont rarement gaussiens: prenons une photo de notre environnement par exemple, on y verra une multitude de discontinuités qui sont essentielles pour distinguer les objets des uns des autres. Et là, les Processus Gaussiens ne sont pas capables d'appréhender par exemple les phénomènes de turbulence, les textures, etc, par contre l'aspect très spectaculaire des réseaux de neurones semblent bien capables de rendre compte de ces phénomènes. Mais dans ce cas, la caractérisation des ensembles typiques est beaucoup plus complexe.

La question centrale développée déjà dans les cours antérieurs, est la suivante: comment, ou par quel mécanisme sous-jacent, les familles de probabilités induites par les réseaux de neurones sont-elles génériques? Au sens que le même type de cascades d'opérateurs (convolution, rectificateurs...) est capable de capturer les caractéristiques de problèmes largement indépendants/déconnectés.

2.7 Complexité, structure des architectures

C'est un thème que S. Mallat a abordé dans son Cours de 2020 [26] à propos du rôle du livre de **Herbert A. Simon** (1916-2001) [27], *The Architecture of Complexity*, paru en 1962 [28]. La question soulevée est: existe-t'il des familles génériques pour le traitement de données?

Herbert A. Simon écrit là un ouvrage totalement différent de ceux de R. Fisher et C. Shannon, où il prend du recul par rapport au domaine. En particulier, il étudie quelles sont les *structures génériques* du "monde" (c'est-à-dire en observant ce qu'il se passe en biologie, en traitement du langage, en physique...):

- la structure est ***hiérarchique*** quasiment tout le temps;
- une ***explication dynamique*** (temporelle) de cette structure hiérarchique est la ***recherche de stabilité*** (survie);
- la ***séparabilité des échelles*** (dans la hiérarchie) qui permet de détruire la malédiction de la dimension;
- la description temporelle doit être vue comme des ***processus agrégatifs*** tendant à la stabilité globale, et non une succession d'états statiques où l'ordre serait établi dés le départ.

Aussi enthousiasmante et inspirante soit la lecture de ce type d'article, *in fine* on reste un peu coi car il n'y aucun modèle mathématique auquel se raccrocher, et l'impact n'est pas du tout du même ordre que les articles de Fisher et Shannon. Cependant, ce qui a changé par rapport au temps où Simon écrit son article, c'est que l'on a des algorithmes qui mettent en œuvre les structures hiérarchiques (ex. l'enchainement de convolutions suivi de sous-échantillonnage permet de changer l'échelle d'analyse d'une image par exemple) mais les mathématiques ne sont pas encore en mesure de tout comprendre. Néanmoins, pour tenter de comprendre les familles de probabilités sous-tendues par les réseaux de neurones, il faut déjà bien comprendre la base des théories de Fisher et Shannon. Ce que l'on va aborder dans la suite à travers la ***Théorie du Codage*** et en particulier le codage d'images.

---

26. Voir note du Cours 2020 Sec. 3.2.

27. prix de la Banque de Suède d'Économie en 1978, mais surtout prix Turing en 1975 pour ses contributions à l'Intelligence Artificielle dont il est l'un des pionniers aux USA avec Allen Newell (1927-92) dont il partage le prix Turing.

28. Proceedings of the American Philosophical Society, Vol. 106, No. 6. (Dec. 12, 1962), pp. 467-482. https://www2.econ.iastate.edu/tesfatsi/ArchitectureOfComplexity.HSimon1962.pdf.

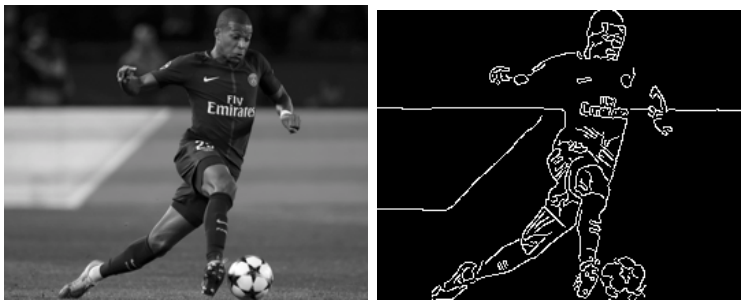

FIGURE 6 – Le dessin des contours permet de bien se représenter ce que l'image nous dit. Et encore l'algorithme utilisé ici donne beaucoup (trop peut-être) d'informations.

2.8 Codage des images

Le codage des images revient d'une certaine manière à spécifier les ensembles typiques. Si on prend les Processus Gaussiens pour décrire une image, nous verrons que cela revient en gros à considérer que les structures sont des fonctions régulières sans aucune discontinuités, contours, etc. Donc, ça n'ira pas, si l'on veut de la finesse dans la description de l'image. Ce qui va nous aider (vraiment) c'est l'utilisation de ***représentations parcimonieuses*** qui était le sujet du Cours de 2021. Nous allons donc faire le lien avec le codage.

Intuitivement, en dessinant les contours des objets dans une image (Fig. 6), on réalise que l'on a déjà une bonne description de celle-ci. En faisant cela, on regarde les lieux des ***singularités***. Peut-on faire une description de l'image en termes de "*transitions*"? Si oui alors on s'aperçoit que l'information des pixels est très redondante, et on peut la compresser. Donc, le schéma est le suivant:

- être capable de représenter les transitions/variations,
- faire cela à différentes résolutions.

On peut dérouler ce schéma en utilisant une ***base orthogonale d'ondelettes***[29]. Les géométries des ensembles typiques sont alors très allongées selon les axes de la base du fait de la ***parcimonie***. Ainsi, dans certaines directions les coefficients de la décomposition en ondelette peuvent être grands mais la plupart des coefficients sont proches de zéro (Fig. 7). Dans le cours de 2021, nous avions vu l'équivalence entre la capacité à faire

29. Voir également les éléments sur le sujet dans les Cours de 2018, 2020 et 2021.

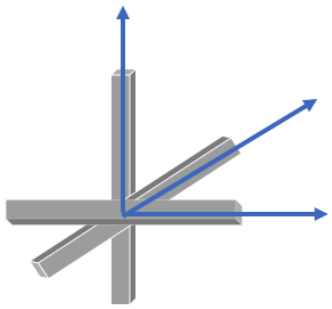

FIGURE 7 – Schématisation des ensembles typiques: quand on utilise une base orthonormée d'ondelettes celle-ci concentre les coefficients à être essentiellement non nul le long des axes de la décomposition.

de l'***Approximation***, le fait qu'il existe des représentations ***Parcimonieuses***, et l'existence d'une ***Régularité*** sous-jacente, triptyque baptisé alors triangle RAP. La mise en oeuvre des codages qui en découle est le standard JPEG-2000 développé dans les années 1997-2000 et officialisé en 2015 par les trois organismes ISO, IEC et l'UIT. Nous verrons qu'il associe les ondelettes à un codage entropique pour compresser les images.

Pour ce qui concerne l'organisation du cours et des challenges, il vaut mieux regarder la vidéo du cours enregistré (>1:15 du début).

3. Séance du 26 Janv.

3.1 Retour sur déterminisme vs probabilisme

S. Mallat revient en début de cours sur la différence d'approche entre *déterminisme* et *stochastique* en grande dimension (Sec. 2.2).

Rappelons nous quand on se demande comment relier une variable y à une autre x, dans l'approche *déterminisme*, on pense à une fonction f inconnue mais qui préexiste, telle que $y = f(x)$. Et si l'on dispose d'observations $\{x_i, y_i\}_{i\leq n}$ (Fig.1) alors on connait les valeurs de la dite fonction aux points $(x_i)_{i\leq n}$. Dans ce cadre le problème mathématique est un problème d'*interpolation* qui se fait bien dès que la fonction est *régulière*. Ce que

l'on a mentionné à la séance précédente, c'est que l'interpolation en grande dimension est potentiellement très difficile à cause de la *malédiction de la dimensionalité*. Néanmoins, gardons à l'esprit qu'***en basse dimension l'interpolation est très efficace***, et pour peu que l'on réussisse à redéfinir le problème en réduisant la dimensionalité, alors on a un très bon outil abondamment utilisé en Physique où $x \in \mathbb{R}, \mathbb{R}^2, \mathbb{R}^3$ voire des problèmes faisant intervenir le temps, mais aussi par exemple en traitement d'images où l'interpolation permet de combler les voies "mortes" d'un capteur CCD. Ce qui se passe en basse dimension, qui est la clé du succès, c'est que la densité des points d'échantillonnage est grande (ou peu l'être), alors qu'en grande dimension le problème devient totalement différent comme on l'a vu. D'où l'idée d'envisager une approche *probabiliste*.

Maintenant, concernant les types de fonctions f, dans le cours de 2021, on s'était pencher sur les notions de *régularité* et la relation entre *approximation* et *parcimonie*. En basse dimension, l'analyse fonctionnelle se pose la question quel est l'espace auquel appartient la dite fonction? ex. dans les espaces de Sobolev f a des dérivées d'un certain ordre, pour les espaces de Hölder f a des singularités d'un certain type, etc. Tout ça se passe bien en basse dimension. Si on se place en grande dimension, ce sur quoi on s'était basé également dans le Cours de 2021, c'est que les x qui nous intéressent *primo* sont indexés par u ($x(u)$) une variable de *basse dimension* (ex. le temps dans un échantillon sonore, la position d'un pixel dans une image) et *secundo* ces x se concentrent dans des zones relativement petites par rapport à la taille de l'espace possible (voir la notion d'*ensemble typique* Sec. 2.5 et Fig. 5).

Dans le cas *probabiliste*, ce qui nous intéresse ce n'est pas tant f mais $p(y|x)$. Et donc, on a affaire à des problèmes d'estimations de probabilités, et la notion fondamentale qui va nous permettre de s'affranchir de la malédiction de la dimensionalité, c'est l'***indépendance***. C'est vraiment le point qui rend effectives les stats.

3.2 La notion d'indépendance et de séparabilité

Soit des observations (par ex. les pixels d'une images, les échantillons sonores d'une trame musicale, les mots d'un texte), $x = \{x_i\}_{i \leq d}$, si elles sont indépendantes alors, on a

$$p(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} p(x_i) \tag{28}$$

Pourquoi est-ce crucial en grande dimension? C'est grâce à un argument déjà évoqué dans le cours de 2020 [30] à savoir la ***séparabilité des variables***. Prenons le logarithme, il vient

$$\log p(x_1, x_2, \ldots, x_d) = \sum_{i=1}^{d} \log p(x_i) \tag{29}$$

c'est-à-dire que d'1 problème à d variables, on se ramène à d problèmes à 1 variable et l'on retombe dans de la basse dimension "classique". Dans le cas *déterministe*, on se poserait la question de savoir s'il est possible d'écrire la fonction $f(x)$ également sous la forme d'une somme de fonctions f_k portant sur un sous-ensemble des variables de x, afin de réduire la dimensionalité de chaque f_k.

Donc si en termes déterministes, on dit *"cherchons à séparer le problème initial de grande dimension en sous problèmes de plus petite dimension, donc plus simples"* (à la manière de R. Descartes du *Discours de la Méthode*), en termes probabilistes on dit *"indépendance des variables aléatoires"*.

Le hic quand on est confronté à de vraies observations, c'est d'essayer de trouver cette "indépendance" si elle existe. Par exemple, prenons une photo d'une écorce d'arbre, et on se pose la question de générer de nouvelles images d'écorces d'arbres. Le problème est que les valeurs des pixels de la photo d'origine ne sont pas du tout indépendantes ou peu-être il y a une corrélation à une certaine échelle et beaucoup moins à une autre, et de nouveau une corrélation à une autre, etc. ***Donc, soit on se donne a priori des observations indépendantes et là tout se passe bien, soit il faut trouver les structures/les échelles qui rendent les variables indépendantes.***

3.3 La loi des grands nombres: convergence vers la moyenne

La loi des grands nombres nous dit que quand on peut disposer de beaucoup d'observations, les fréquences convergent vers des espérances, et l'on observe des phénomènes moyens. Et là encore en sous-jacent, on a la notion d'indépendance [31]. La base mathé-

30. Voir Cours 2020 Sec. 4.3

31. NDJE: cette notion est très importante pour juger de l'efficacité de certaines méthodes statistiques que sont la génération de chaine de Markov. On définit alors l'efficacité de sampling ou bien la taille de l'ensemble d'échantillons indépendants, pour juger de la fiabilité des statistiques que l'on peut en tirer comme des intervalles de confiance.

matique est constituée par les travaux de R. Fisher de 1922 et les notions de *consistance d'estimateur*, de *maximum de vraisemblance*, d'*information* et de bornes/limites à l'*approximation*. S. Mallat nous dit que par rapport au déroulé classique d'un cours de statistique, il fera des incises dans le domaine de la grande dimension pour montrer ***le caractère non-évident de ces notions*** et que derrière il y a la notion d'***optimisation***. On peut voir les problèmes sous deux points de vue: soit celui des statistiques, soit celui de l'optimisation. Par exemple le ***Hessien*** de la vraisemblance va nous permettre de contrôler la convergence, et l'erreur des estimateurs laquelle est reliée à l'information de Fisher.

Concernant la convergence d'une série de n variables aléatoires, il y a celle introduite par Andrey N. Kolmogorov [32] (1903-87) qui définit la ***la loi forte des grands nombres*** que l'on peut résumer selon l'expression: si on a une variable aléatoire $(v.a)$ qui dépend de n, le nombre d'observations, telle que $A_n \xrightarrow[n\to\infty]{} A$

$$\mathbb{P}\left[\lim_{n\to\infty} A_n = A\right] = 1 \tag{30}$$

mais celle que l'on utilise généralement, c'est plutôt ***la loi faible des grands nombres*** qui stipule que

$$\boxed{\left(\forall \varepsilon > 0,\ \lim_{n\to\infty} \mathbb{P}\left[|A_n - A| \leq \varepsilon\right] = 1\right) \Leftrightarrow \left(A_n \xrightarrow[n\to\infty]{prob.} A\right)} \tag{31}$$

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

Voici le théorème dans le cas où la $v.a$ A_n est la moyenne d'ensemble de n $v.a$ iid:

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

$$\bar{X}_n \xrightarrow[n\to\infty]{prob.} \mu \tag{32}$$

32. 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.

La démonstration de ce théorème est un peu technique dans le cas général, mais surtout son principal défaut est de ne pas nous renseigner sur *la vitesse de convergence*. Donc, on va démontrer ce théorème dans le cas où ***la variance*** σ^2 ***existe*** c'est-à-dire $\mathbb{E}[X_i^2] < \infty$.

Démonstration 1.
Soit la variance de la moyenne empirique $\sigma^2(\bar{X}_n)$, alors selon l'hypothèse de l'indépendance des $v.a$, on a simplement

$$\sigma^2(\bar{X}_n) = \frac{\sigma^2}{n} \tag{33}$$

Donc on comprend bien que plus n augmente, plus $\bar{X}_n$ sera piquée autour de sa moyenne donc de μ, et simultanément les queues de distribution seront faibles. Le point technique ici c'est l'inégalité de Bienaymé-Tchebychev [33] qui est un résultat de concentration de probabilité [34], si $Var[X] = \sigma^2$:

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \alpha\right] \leq \frac{\sigma^2}{\alpha^2} \tag{34}$$

Donc, en combinant les deux résultats, on a

$$\boxed{\mathbb{P}\left[\left|\bar{X}_n - \mu\right| \geq \varepsilon\right] \leq \frac{\sigma^2}{n\varepsilon^2} \xrightarrow[n\to\infty]{} 0} \tag{35}$$

L'importante conséquence est que la *convergence* de $\bar{X}_n$ vers l'espérance μ est en $1/n$ donc *assez rapide*. ■

3.4 Consistance: l'estimation de paramètres

Suivant l'idée de R. Fisher, on se donne n observations et l'on veut estimer la loi de probabilité sous-jacente, en se donnant une famille de probabilités paramétrées $p_\theta(x_i)$. Il

33. NDJE: il s'agit de Irénée-Jules Bienaymé (1796-1878) et de Pafnouti Lvovitch Tchebychev (1821-94), et cette inégalité n'est pas à confondre avec l'inégalité de Tchebychev sur les sommes.

34. La démonstration tient au fait que $\forall x \in \mathbb{R}$, $\mathbf{1}[|x| \geq 1] \leq x^2$ ($\mathbf{1}$: indicatrice), donc appliqué à $(X - \mu)/\alpha$ ($\alpha > 0$) et en se souvenant de la croissance de l'espérance et que $\mathbb{E}[\mathbf{1}[A]] = \mathbb{P}[A]$, on trouve bien inégalité mentionnée.

nous faut alors estimer le "meilleur" θ.

Définition 1 *(**Consistance**)*
Soit une statistique T_n comme fonction de $(X_1, \dots, X_n)$ (n v.a), on dit que c'est un estimateur consistant de θ si

$$T_n(X_1, \dots, X_n) \xrightarrow[n\to\infty]{prob.} \theta \tag{36}$$

Par exemple, un estimateur de l'espérance μ est la moyenne empirique $\bar{X}_n$, et concernant la variance on peut penser à

$$T_n(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 \tag{37}$$

Pour étudier la convergence, il nous faut un petit théorème

Théorème 2 *Soit une séries de v.a A_n qui convergent vers A en probabilité, et soit une fonction continue g, alors $g(A_n)$ converge en probabilité vers $g(A)$.*

Démonstration 2. La démonstration découle de l'hypothèse de continuité qui stipule que

$$\forall \varepsilon > 0 \ \exists \alpha > 0 \ tq. \ |a - a'| \leq \alpha \Rightarrow |g(a) - g(a')| \leq \varepsilon \tag{38}$$

Donc on a

$$1 \geq \mathbb{P}(|g(a) - g(a')| \leq \varepsilon) \geq \mathbb{P}(|a - a'| \leq \alpha) \xrightarrow[n\to\infty]{prob.} 1 \tag{39}$$

d'où le résultat. ■

Donc, comme $\bar{X}_n$ converge en probabilité vers μ, alors $\bar{X}_n^2$ converge également en probabilité vers μ^2. De même en posant $Y_i = X_i^2$, $\bar{Y}_n$ converge vers l'espérance $\mathbb{E}[X_i^2]$ donc

$$T_n(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \xrightarrow[n\to\infty]{prob.} \mathbb{E}[X_i^2] - \mu^2 \tag{40}$$

qui donne la convergence en probabilité vers la variance des X_i. Ce qui va donner la

dynamique de la convergence c'est maintenant la variance des X_i^2. Donc, on va imposer que $\mathbf{E}[X_i^4] < \infty$ pour pouvoir appliquer la loi des grands nombres.

Maintenant dans le cas plus général, R. Fisher se sert du maximum de vraisemblance pour obtenir des estimateurs consistants.

3.5 Maximum de vraisemblance

Tout d'abord donnons nous une définition de la vraisemblance selon Fisher

Définition 2 *(**Vraisemblance**)*
Soit les v.a $X = \{X_i\}_{i\leq n}$ iid, la vraisemblance (likelihood) de ces observations pour un paramètre θ est définie par

$$\mathcal{L}_\theta(X) = p_\theta(X) = \prod_{i=1}^{n} p_\theta(X_i) \tag{41}$$

L'idée de Fisher est de dire alors que si $\mathcal{L}_{\theta_1}(X) > \mathcal{L}_{\theta_2}(X)$ alors la génération des observations est plus probable si on prend $\theta = \theta_1$ que si on prend $\theta = \theta_2$. Par un abus de langage, on opère un raccourcit et on dit souvent "θ_1 est plus probable que θ_2". Il vaudrait mieux dire "θ_1 est un meilleur estimateur que θ_2".

Définition 3 *(**MLE**/maximum likelihood estimator)*
Le MLE est défini comme

$$\hat{\theta}_{MLE}(X) = \underset{\theta}{\operatorname{argmax}}\ \mathcal{L}_\theta(X) \tag{42}$$

La question est de savoir si $\hat{\theta}_{MLE}(X)$ converge vers la vraie [35] valeur de θ. On a besoin de propriétés de régularité des probabilités.

35. NDJE: On a comme hypothèse que les données ont été "générées" effectivement selon une probabilité de la même famille $p_{\theta_{true}}(X)$ que celle avec laquelle on les analyses. Quand on fait des simulations numériques on peut tout contrôler mais dans la vraie vie qu'en est-il si on se trompe de famille?

Propriété 1 *(**régularités**)*
On va supposer les propriétés suivantes sachant que $\theta \in \Omega \subset \mathbb{R}^d$

- *la propriété d'identification*

$$\theta = \theta' \Rightarrow p_\theta = p_{\theta'} \tag{43}$$

- *les supports des* p_θ *sont identiques (pas forcément nécessaire mais pratique car évite les singularités quand on calcule les log-prob)*
- *Les observations sont effectivement générées par un* $\theta^* \in \Omega$

Ainsi, on va pouvoir formaliser l'intuition que l'on a sur le maximum de vraisemblance (Voir aussi Sec. 4.7).

Théorème 3
Soit θ^* *le paramètre la probabilité sous-jacente aux observations* $X = \{X_i\}_{i \leq n}$ *iid, alors*

$$\forall \theta \neq \theta^* \qquad \mathbb{P}\left(\mathcal{L}_{\theta^*}(X) > \mathcal{L}_\theta(X)\right) \xrightarrow[n\to\infty]{prob.} 1 \tag{44}$$

Démonstration 3.
On forme le rapport des likelihoods et en utilisant le log-likehood $\ell(\theta) = \log \mathcal{L}_\theta$ il vient

$$\frac{\ell(\theta)}{\ell(\theta^*)} = \sum_{i=1}^{n} \log \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)} \tag{45}$$

Il nous faut alors évaluer la probabilité pour que $\ell(\theta)/\ell(\theta^*) < 0$, ou d'une manière équivalente

$$\frac{1}{n} \sum_{i=1}^{n} \log \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)} < 0 \tag{46}$$

Si on pose $Y_i = \log \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)}$ on a bien des v.a indépendantes, donc en probabilité le membre de gauche converge vers une espérance, et donc il nous faut évaluer la probabilité de

$$\mathbb{E}_X \left[\log \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)} \right] < 0 \tag{47}$$

Or, le logarithme est une fonction concave, et l'inégalité de Jensen nous donne alors que

$$\phi \text{ fnt. concave} \Rightarrow \phi(\mathbb{E}(X)) \geq \mathbb{E}(\phi(X)) \tag{48}$$

la stricte concavité induit une inégalité stricte. Ainsi, on sait que

$$\mathbb{E}_X\left[\log \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)}\right] < \log\left(\mathbb{E}_X\left[\frac{p_\theta(X_i)}{p_{\theta^*}(X_i)}\right]\right) \tag{49}$$

Or, les observables sont tirées selon la loi $p_{\theta^*}(X)$, donc

$$\mathbb{E}_X\left[\frac{p_\theta(X_i)}{p_{\theta^*}(X_i)}\right] = \int p_{\theta^*}(x)\frac{p_\theta(x)}{p_{\theta^*}(x)}dx = 1 \tag{50}$$

Ainsi, on a une convergence en probabilité telle que

$$\exists \mu \quad tq. \quad \bar{Y_n} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)} \xrightarrow[n\to\infty]{prob.} \mu \ < 0 \tag{51}$$

Alors si on choisit $\varepsilon = |\mu|/2$, on garantit que $\bar{Y_n}$ est négatif car selon la loi des grands nombres (Th. 1)

$$\mathbb{P}(|\bar{Y_n} - \mu| \leq \varepsilon) \xrightarrow[n\to\infty]{prob.} 1 \tag{52}$$

Donc en revenant à la question posée (Eq. 46), on garantit que $\ell(\theta) < \ell(\theta^*)$ avec une probabilité qui tend vers 1 quand n tend vers l'infini, ce qui donne le théorème. ■

3.6 Quelques exemples

Nous allons voir à travers quelques exemples que les notions décrites dans les sections précédente ne sont pas si triviales que cela.

3.6.1 Estimateur médian vs moyenne empirique

Donc, on veut déterminer $\hat{\theta}$ qui réalise le maximum du likelihood $\mathcal{L}(\theta) = p_\theta(x)$ ou plutôt le log-likelihood noté $\ell(\theta)$ (Déf. 3). Dans ce contexte, si l'on pose

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\ \ell(\theta) \Rightarrow \left.\frac{\partial \ell}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0 \tag{53}$$

on définit le ***score*** qui n'est autre que la dérivée de $\ell(\theta)$ que l'on va donc tenter d'annuler.

Prenons la distribution de Laplace [36]

$$p_\theta(x) = \frac{1}{2}\exp\{-|x-\theta|\} \tag{54}$$

on veut en identifier le paramètre θ, donc déroulons le formalisme. Si on dispose de n observables

$$\ell(\theta) = -n\log 2 - \sum_{i=1}^{n}|x_i-\theta| \Rightarrow \partial_\theta \ell(\theta) = \sum_{i=1}^{n}\text{sign}(x_i-\theta) \tag{55}$$

Pour annuler le score il faut qu'il y ait autant de signes positifs que de signes négatifs et donc

$$\hat{\theta}_{Laplace} = \text{median}(\{x_i\}_{i\leq n}) \tag{56}$$

Si nous avions pris une loi gaussienne de *variance connue* alors

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\} \tag{57}$$

et donc

$$\ell(\theta) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_i (x_i-\theta)^2 \Rightarrow \partial_\theta\ell(\theta) \propto \sum_i(\theta - x_i) = n\theta - \sum_i x_i \tag{58}$$

Ainsi, l'estimateur est la moyenne empirique des x_i

$$\hat{\theta}_{Gauss} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{59}$$

Ce qui est curieux c'est que dans les deux cas, Laplace vs Gauss, on doit estimer la moyenne

36. Ici on prend le second paramètre de la distribution de Laplace égal à 1.

de la distribution et pourtant on a deux estimateurs, le premier c'est le *médian*, le second c'est la *moyenne empirique*. Le problème de la distribution de Laplace réside dans la lente décroissante des queues de distributions qui peuvent générer des observations à grande distance de la moyenne (on dit des *outliers*). Or, si l'on calcule une moyenne empirique avec des outliers qui apparaissent à faible fréquence, on a de grandes dispersions, alors que le calcul du médian est beaucoup plus robuste contre les outliers. Ce phénomène d'ouliers n'est pas juste une anecdote de mathématiciens car en traitement du signal, en physique, en économie, etc, on est confronté à ce genre de problématique.

3.6.2 Descente de gradients en grande dimension

Maintenant passons à la grande dimension et tout d'abord étudions le ***classificateur logistique***[37]. C'est une classification et pourtant on l'appelle souvent *régression logistique*. On se place donc dans le cas où l'on veut estimer la probabilité conditionnelle $p(y|x)$ (y une étiquette/un label, et x une observation) dans le formalisme de Fisher où la famille de probabilité est indexée par θ. Donc, on veut identifier le meilleur y ($\hat{y}$) pour x donné, mais avant cela il nous faut déterminer le meilleur θ à partir d'un lot d'entrainement $\{x_i, y_i\}_{i\leq n}$. La famille de probabilité est définie selon

$$p_\theta(y|x) = \frac{e^{\langle x,\theta_y\rangle}}{\sum_{y'} e^{\langle x,\theta_{y'}\rangle}} = \text{softmax}(\langle x, \theta_y\rangle) \tag{60}$$

avec $x \in \mathbb{R}^d$, et l'objectif est de trouver les directions privilégiées $\hat{\theta}_y$ qui pointent vers les zones où les observations ayant les mêmes labels s'agrègent (Fig. 8), car alors on choisira comme estimateur de la classe pour une nouvelle observation x^{new} le label tel que

$$\hat{y} = \underset{y}{\text{argmax}}\ p_{\hat{\theta}}(y|x^{new}) \tag{61}$$

On va faire entrer ce problème dans celui de l'***identification des lois exponentielles*** qui couvre toute la Physique Statistique.

Définissons des ***hot-vectors*** (dim. $K \times 1$) très utilisés en machine learning:

$$y_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T \tag{62}$$

37. Cours 2018 Sec. 9.6, Cours 2019 Sec. 7.3.3

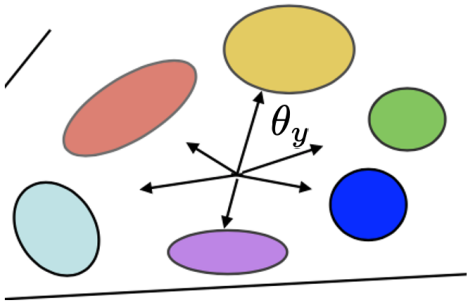

FIGURE 8 – Schématisation de l'objectif d'une classification logistique qui tente de trouver les directions θ_y pointant vers les différentes zones d'agrégats d'observations ayant le même label.

où le 1 est positionné à la i-eme place pour identifier la classe $c(x_i) = c_i$ de l'observation x_i parmi les K classes possibles, et en même temps Θ est une matrice $d \times K$ définie par les vecteurs colonnes θ_y mentionnés ci-dessus telle que

$$\Theta = (\theta_1, \theta_2, \dots, \theta_K) \tag{63}$$

Avec ces notations

$$\langle x, \theta_{c_i} \rangle = \langle x, \Theta\ y_i \rangle = x^T \Theta\ y_i \tag{64}$$

Donc, on va se placer dans la famille des probabilités suivantes

$$p_\theta(y|x) = Z_\Theta^{-1}(x) \exp\Big\{x^T \Theta y\Big\} \tag{65}$$

Le point important est que ***l'argument de l'exponentiel est linéaire en les paramètres***, ce qui va nous facilité la tâche. Le log-likelihood pour n observations [38] $(x_i, y_i)_{i \le n}$ devient

$$\ell(\Theta) = \sum_{i=1}^{n} x_i^T \Theta\ y_i - \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \exp\Big\{x_i^T \Theta\ y_k\Big\}\right) = -\tilde{\ell}(\Theta) \tag{66}$$

Notons que les termes $y_k x_i^T$ forment une matrice qui représente la corrélation entre les

38. Attention: ici i indice une observation, et y_i est le hot-vecteur associé à cette observation qui encode la classe de celle-ci.

observations et les classes.

$$x^T\Theta\ y = \sum_{kk'} x_{1,k}\Theta_{k,k'}y_{k',1} = \sum_{kk'}\Theta_{k,k'}(yx^T)_{k',k} := \Theta \bullet (yx^T) \tag{67}$$

La notation $\bullet$ signifie que l'on aplatit (*flatten*) les coefficients des matrices pour en constituer un vecteur de dimension Kd et faire apparaitre un produit scalaire.

Comment obtient-on les coefficients de Θ (problème d'optimisation)? Dans le cas qui nous occupe, on va procéder à *une descente de gradient*[39] (GD) en prenant la fonction de coût[40] $\tilde{\ell}(\Theta) = -\ell(\Theta)$, et l'on sait que la méthode *converge*. En sous-jacent, pour que cela converge, il y a une ***propriété de convexité***. L'algorithme GD procède par une initialisation des paramètres Θ_0, et étape après étape opère une actualisation selon (t peut être vu comme un temps discret)

$$\Theta_t = \Theta_{t-1} - \eta\nabla_\Theta\tilde{\ell}(\Theta_{t-1}) \tag{68}$$

avec $\eta > 0$. Soit H la matrice Hessienne

$$H[\tilde{\ell}][\Theta] = \left(\frac{\partial^2\tilde{\ell}}{\partial\Theta_i\partial\Theta_j}\right) \tag{69}$$

Comme H est une matrice symétrique, on peut la diagonaliser et si toutes les v.p sont positives, on dira que H est positive, notée $H \geq 0$, alors la fonction est convexe. En 1D, cela correspond à la courbure par ex. de la fonction x^2. Si on effectue un développement autour de Θ_0, alors ($g = \nabla_\Theta\tilde{\ell}(\Theta_0)$)

$$\begin{aligned}\tilde{\ell}(\Theta) &= \tilde{\ell}(\Theta_0) + (\Theta-\Theta_0)^T\nabla_\Theta\tilde{\ell}(\Theta_0) + \frac{1}{2}(\Theta-\Theta_0)^T H[\tilde{\ell}](\Theta_0)(\Theta-\Theta_0)\\ &= \tilde{\ell}(\Theta_0) - \eta\|g\|^2 + \frac{\eta^2}{2}g^T H[\tilde{\ell}](\Theta_0)g\end{aligned} \tag{70}$$

La méthode GD fait bien diminuer $\tilde{\ell}(\Theta)$ au premier ordre et la ***condition d'existence d'un minimum*** implique que

$$\|g\| = 0, \qquad g^T H[\tilde{\ell}](\Theta_0)g \geq 0 \tag{71}$$

39. Voir Cours 2018 Sec. 10 et 2019 Sec. 8 par exemple.

40. NDJE: on utilise le plus souvent la descente du gradient car la majorité des softwares se place dans cette optique.

Le pas optimal est alors ($\nabla_\Theta \tilde{\ell}(\Theta_1) = 0$)

$$\Theta_1 - \Theta_0 = -\left(H[\tilde{\ell}](\Theta_0)\right)^{-1} \nabla_\Theta \tilde{\ell}(\Theta_0) \tag{72}$$

Mais pour que cela se passe bien il faut que le Hessien soit inversible (**la plus petite valeur propre du Hessien soit non nulle**). Ce schéma est ***du 2nd ordre*** où l'on peut disposer des dérivées secondes. Le problème c'est qu'en grande dimension, ce calcul est très couteux voire impossible à faire, on a donc recours alors à un schéma du **1er ordre** (Eq. 68) avec différentes stratégies dites de scheduling pour faire évoluer le paramètre (learning rate) η en fonction du temps (t). En particulier, le facteur η est borné, car on ne veut pas faire un pas plus grand que ce que la méthode avec le Hessien nous autorise. Soit donc λ_{max} la plus grande valeur propre du Hessien, on a la borne suivante

$$\eta < \frac{1}{\lambda_{max}} \tag{73}$$

Cependant, si l'écart entre la plus petite valeur propre du Hessien, notée λ_{min} et la plus grande λ_{max} est trop important alors en forçant de trop petits steps pour contraindre la direction associée à λ_{max}, et on va se retrouver coincé à stagner dans la direction associée à λ_{min}. Ceci se traduit par la notion de conditionnement du Hessien:

Théorème 4 *(Convergence de la méthode GD)*
Soit $\lambda_{min} > 0$ et λ_{max} les valeurs minimales et maximales du propres du ***Hessien****, la descente de gradient converge si $\eta \leq \frac{1}{\lambda_{max}}$ et l'écart entre Θ_t et la valeur optimale Θ^* est donné par*

$$\|\Theta_t - \Theta^*\| \leq \left(1 - \frac{\lambda_{min}}{\lambda_{max}}\right)^t \|\Theta_0 - \Theta^*\| \leq \|\Theta_0 - \Theta^*\| \exp\left\{-\frac{\lambda_{min}}{\lambda_{max}} t\right\} \tag{74}$$

Le taux de conditionnement est donné par $\tau = \frac{\lambda_{min}}{\lambda_{max}}$ (τ est l'inverse du conditionnement du Hessien).

Le théorème nous dit que ***la méthode de descente de gradient converge surtout dans le cas des familles exponentielles (linéaires dans les paramètres) mais, la convergence peut être très lente si le Hessien est mal conditionné***. Or, ceci est très important, et toute la problématique du conditionnement du Hessien, c'est ***l'information de Fisher*** qui

définit les propriétés statistiques de l'estimateur. On voit par ce biais la jonction entre le domaine de l'*optimisation* et celui des *statistiques* au cœur du Machine Learning. On ne peut pas penser l'un sans l'autre.

4. Séance du 2 Févr.

4.1 Petit retour sur la séance précédente

Nous avons abordé le lien entre *optimisation* et *statistiques* qui sont deux domaines indissociables du Machine Learning actuel. Nous avions commencé à étudier l'élaboration de la théorie de Fisher du maximum de vraisemblance, à travers le cas de la *famille exponentielle* des probabilités, où le log-prob dépend linéairement des paramètres. Ce cas recouvre quasiment toute la Physique Statistique. Nous allons poursuivre cette étude car, certes les mathématiques sont un peu plus simples, les algorithmes convergent vers le minimum unique, mais néanmoins cela nous permet d'aborder toute la problématique de la grande dimension. Et nous allons étudier *la convergence*, *la consistance* des estimateurs de maximum de vraisemblance qui nous mènera la fois prochaine à la notion d'***Information de Fisher*** laquelle à travers le Hessien régule les *conditions de convergence* des algorithmes et définit la *géométrie de l'espace d'optimisation*, et les *erreurs sur l'estimation des paramètres* (borne de Cramér-Rao).

En se reportant à la section 3.6.2, l'algorithme de *descente* de gradient sur $\tilde{\ell}(\Theta)$ peut être aussi vu comme une *montée* de gradient sur le log-prob. $\ell(\Theta)$. Donc, pour mémoire [41] (Eqs. 68,69) à l'étape t de l'algorithme, la mise à jour des paramètres Θ se fait via la relation suivante

$$\Theta_t - \Theta_{t-1} = -\eta \nabla_\Theta \tilde{\ell}(\Theta_{t-1}) = \eta \nabla_\Theta \ell(\Theta_{t-1}) \tag{75}$$

avec $\eta > 0$; et la matrice Hessienne

$$H[\ell][\Theta] = -\left(\frac{\partial^2 \ell}{\partial \Theta_i \partial \Theta_j}\right) = -H[\tilde{\ell}][\Theta] \tag{76}$$

41. NDJE: je reste cohérent avec mes notations de la séance précédente. Dans la vidéo S. Mallat utilise la notation θ pour les paramètres. J'espère que cela n'est pas trop déroutant.

doit être *positive* pour que la minimisation soit *convexe*. Si on requiert une *stricte positivité*, c'est-à-dire que la plus petite valeur propre de H est non nulle [42], alors la convergence est garantie mais la vitesse de convergence peut être très lente. Il faut regarder ***le conditionnement du Hessien*** (κ) ou le *tau de conditionnement* τ

$$\tau = \frac{\lambda_{min}}{\lambda_{max}} = \kappa^{-1} \tag{77}$$

ce qui nous à mener au théorème 4 stipulant que

$$\|\Theta_t - \Theta^*\| \leq \|\Theta_0 - \Theta^*\| e^{-t/\kappa} \tag{78}$$

Le conditionnement est d'autant meilleur que $\kappa \approx 1$. Si par contre $\kappa \gg 1$ la convergence est très lente. Cela se passe par exemple dans le cas de la figure 9 où une des deux directions dans le plan des paramètres a une courbure très faible. En effet le pas η est contrôlé par la plus grande courbure pour ne pas faire de sauts trop grands. Le remède à cela est de faire dépendre le pas η de la direction (méthode du second ordre où $\eta = H^{-1}(\Theta_t)$) pour s'ajuster au mieux, et augmenter la vitesse de convergence. Cependant, *en grande dimension on ne peut jamais utiliser une méthode de second ordre* car le hessien est une matrice énorme qu'il est impossible d'estimer et *a fortiori* d'inverser. Il faudra d'autres méthodes pour faire l'apprentissage des grands réseaux de neurones. Mais, néanmoins on peut tenter de pré-conditionner le Hessien.

4.2 Cas des distributions exponentielles

Ces distributions qui couvrent la Physique Statistique, couvrent également le cas de la classification logistique abordée la séance précédente. Prenons l'expression de la famille de probabilité selon

$$p_\theta(x) = Z_\theta^{-1} e^{-\theta \bullet U(x)} \tag{79}$$

où le symbole $\bullet$ a été exagéré volontairement pour préciser qu'il s'agit d'un produit scalaire potentiellement en grande dimension (par la suite, il sera réduit à un . voir disparaitra), et $U(x)$ qui en Physique est le potentiel, est une famille de fonctions $\{U_k(x)\}_{k \leq p}$ représentant

42. rappel: toutes les valeurs propres du Hessien sont positives on nulles.

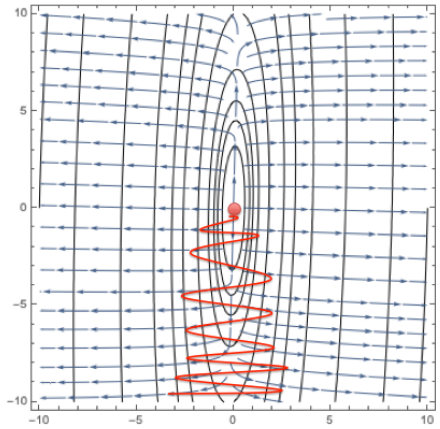

FIGURE 9 – Cas certes convexe mais où le paysage n'est pas favorable car il y a une direction qui a une très faible courbure: l'algorithme "oscille" dans la direction de forte courbure avec une progression très lente dans celle de faible courbure.

par exemple différents types d'interactions, et pour être complet

$$\theta \bullet U(x) = \sum_k \theta_k U_k(x) \tag{80}$$

Concernant la constante Z_θ (*fonction de partition* en Phys. Stat), elle est telle que

$$\int p_\theta(x)dx = 1 \Rightarrow Z_\theta = \int e^{-\theta \bullet U(x)}\, dx \tag{81}$$

On supposera les conditions réunies pour que cette intégrale ait un sens: typiquement en Physique les potentiels d'interactions sont soit à portée finie, soit évanescent à l'infini. Cette famille de probabilités rend le log-vraisemblance simple à calculer

$$-\ell(\theta) = -\log p_\theta(x) = \log Z_\theta + \theta \bullet U(x) \tag{82}$$

Remarquons que même si les $U_k(x)$ sont potentiellement non-linéaires en x, ce qui importe dans la méthode d'optimisation, c'est le gradient de $-\ell(\theta)$ par rapport à θ et non par rapport x.

Dans le cas d'un réseau de neurones, la fonction $U(x)$ est notée $\Phi(x)$ sur la figure 4, et les paramètres θ permettent de construire un estimateur de la log-proba. Le $U(x)$

est donc le résultat de l'enchainement des opérateurs linéaires et non-linéaires à travers desquels passe l'entrée x. Notons cependant que dans le cas des réseaux de neurones, $U(x)$ dépend elle-même de paramètres. Mais, on suppose que l'on a suffisamment d'information *a priori*[43] (ex. symétries du système) pour ne pas avoir à apprendre $U(x)$.

Calculons à présent toutes les quantités importantes qui vont nous permettre d'explorer les notions générales de l'optimisation.

Théorème 5

Soit donc la fonction de partition

$$Z_\theta = \int e^{-\theta \bullet U(x)}\, dx \tag{83}$$

elle permet de calculer toutes les quantités "moyennes"[a]

$$-\nabla_\theta \log Z_\theta = \mathop{\mathbb{E}}_{x\sim p_\theta(x)}[U] \tag{84}$$

Concernant la minimisation (descente de gradient), nous avons

$$-\nabla \ell(\theta) = U(x) - \mathbb{E}_\theta[U] \tag{85}$$

Dans le cas d'une réalisation des observables x, le jeu de paramètres vers lequel la minimisation aboutit satisfait $U(x) = \mathbb{E}_\theta[U]$. Enfin, le Hessien qui gouverne la vitesse de convergence est donné par la covariance du potentiel U, à savoir que

$$-H[\ell](\theta) = Cov_\theta(U) \tag{86}$$

a. Pour simplifier la notation, on notera $\mathbb{E}_\theta$ l'espérance du membre de droite.

Démonstration 5.

Pour démontrer les deux premiers résultats, il suffit de calculer les gradients ce qui devient élémentaire pour la famille exponentielle envisagée ici. Notons en passant que si l'on

43. Thématique du Cours de 2020, en particulier voir Sec. 9.5 *Opérateurs de Scattering*.

considère la dépendance par rapport à 1 paramètre particulier θ_k, on a

$$-\nabla_{\theta_k}\ell(\theta) = U_k(x) - Z_\theta^{-1}\int U_k(x)e^{-\theta\bullet U(x)}\,dx = U_k(x) - \mathbb{E}_\theta[U_k] \tag{87}$$

que l'on peut vectoriser facilement. Concernant le Hessien, comme les potentiels $U_k(x)$ ne dépendent pas des paramètres θ, il vient

$$\begin{aligned}-\nabla_{\theta_q}\nabla_{\theta_k}\ell(\theta) &= -Z_\theta^{-1}\int U_q(x)e^{-\theta\bullet U(x)}\,dx \times Z_\theta^{-1}\int U_k(x)e^{-\theta\bullet U(x)}\,dx \\ &\quad + Z_\theta^{-1}\int U_q(x)U_k(x)e^{-\theta\bullet U(x)}\,dx \\ &= -\mathbb{E}_\theta[U_q]\mathbb{E}_\theta[U_k] + \mathbb{E}_\theta[U_qU_k] = cov_\theta(U_q,U_k)\end{aligned} \tag{88}$$

que l'on peut également mettre sous forme matricielle si l'on considère U comme un vecteur $p\times 1$ et UU^T donc de dimension $p\times p$:

$$-H[\ell](\theta) = \mathbb{E}_\theta[UU^T] - \mathbb{E}_\theta[U]\mathbb{E}_\theta[U^T] \tag{89}$$

■

Il est tout à fait remarquable que $H[\ell](\theta)$ ne dépend pas de x: en effet il ne dépend que des espérances qui sont des moyennes en probabilité où l'on intègre sur x.

4.3 La consistance (BatchNorm)

On aimerait que le Hessien soit aussi proche que possible de l'identité afin d'assurer un conditionnement optimal. Supposons que $\mathbb{E}_\theta[U] = 0$ les termes diagonaux du Hessien sont les variances $\sigma_k^2 = \mathbb{E}_\theta[U_k^2]$. Alors, on peut procéder à un rescaling

$$U_k' = \frac{U_k}{\sigma_k} \tag{90}$$

ce qui impose les termes diagonaux du nouveau Hessien à être égaux à 1, et de ce fait tend à rendre le conditionnement meilleur, et accélère ainsi l'optimisation. L'opération qui fait cela dans les réseaux de neurones est la ***BatchNorm***[44].

44. Voir Cours 2019 Sec. 8.2.3.

Est-ce suffisant d'imposer que les éléments de la diagonale soit égaux à 1? Prenons un contre-exemple, en utilisant l'opérateur (discret) de la dérivée seconde

$$-f''(x) \approx \frac{-f(x-h) + 2f(x) - f(x+h)}{h^2} \tag{91}$$

Soit par exemple la matrice bande suivante

$$O = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 & -1 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ & & & \vdots & & \\ -1 & 0 & \dots & 0 & -1 & 2 \end{pmatrix} \tag{92}$$

La remarque (fondamentale) à faire par réflexe, c'est que O est un opérateur de convolution, donc diagonalisable sur une base de Fourier [45]. Soit on fait l'analyse en discret et on se rend compte que les valeurs propres extrêmes de la matrice sont proches de 0 et de 4, soit on fait l'analyse dans le continu, les vecteurs propres de l'opérateur dérivée seconde (au signe près) sont de type $e^{i\omega t}$ et les valeurs propres sont ω^2, donc on se retrouve avec un conditionnement également très mauvais. Tout ça pour dire que "conditionner" la diagonale du Hessien ne suffit pas. ***Il faut/faudrait se placer dans une représentation où la base naturelle est celle de Fourier***, car alors la BatchNorm garantit que les termes diagonaux sont égaux à 1 sans termes de bandes. En fait, l'usage de la BatchNorm revient dans ces conditions à utiliser une technique du second ordre sans le dire. Le hic, c'est que l'on ne connait pas *a priori* qu'elle est la base qui diagonalise la représentation $U(x)$ surtout quand il y a des non-linéarités. On aimerait s'en rapprocher néanmoins et c'est le cœur de la ***construction des architectures des réseaux de neurones***.

4.4 Lien avec la géométrie de l'Information

NDJE: S. Mallat mentionne les 2 séminaires dédiés à ce sujet dont le premier après cette séance, et l'autre associé à la séance de la fois prochaine.

45. Voir par exemple Cours 2021 Sec. 3.4 *Analyse de Fourier*, Cours 2020 Sec. 6.2, Cours 2018 Sec. 5.2 pour un développement de l'Analyse de Fourier en discret. Ainsi que les chapitres du livre de S. Mallat.

L'idée c'est que les $p_\theta(x)$ sont des applications de $\mathbb{R}^p$ dans $\mathbb{R}$, elles formes des variétés et la descente de gradient nous fait évoluer sur cette variété jusqu'à aboutir au point $p_{\theta^*}(x)$. On adjoint à ces variétés des mesures (riemanniennes) qui consistent à prendre en chaque point θ le plan tangent, dont les axes principaux sont définis précisément par le Hessien. Alors se déplacer efficacement sur ces variétés revient à utiliser une méthode du second ordre en utilisant l'inverse du Hessien. On peut voir cela avec la pseudo-distance de Kullback-Leibler qui sera abordée plus loin dans le cours de cette année.

Dans ce qui suit nous allons aborder des exemples et commençons par les distributions gaussiennes.

4.5 Les distributions gaussiennes

Prenons la paramétrisation suivante de la densité de probabilité (moyenne nulle):

$$p(x) = Z^{-1} \exp\left\{-\frac{1}{2} x^T C^{-1} x\right\} \tag{93}$$

avec $Z = (2\pi)^{p/2}|C|^{1/2}$. Dans ce cas le vecteur de paramètres θ est constitué par la matrice de covariance C^{-1}. Pour fixer les notations, x est un vecteur $p \times 1$, et C^{-1} une matrice sym. def. positive de dimension $p \times p$, $[C^{-1}]_{kk'} = c_{kk'}$, et la matrice de Gram $[xx^T]_{kk'} = x_k x_{k'}$:

$$x^T C^{-1} x = \sum_{k,k'} x_k c_{kk'} x_{k'} := C^{-1} \bullet (xx^T) \tag{94}$$

où l'on regroupe les éléments de la matrice de covariance et la matrice de Gram dans deux vecteurs de dimension p^2 pour en calculer un produit scalaire. Ainsi, dans ces conditions, on peut réécrire la densité de probabilité selon le schéma

$$p_\theta(x) = Z_\theta^{-1} \exp\{-\theta \bullet U(x)\} \qquad U(x) = \frac{1}{2} xx^T \tag{95}$$

Rappelons que la matrice de covariance C satisfait pour une réalisation de x ($x \sim p_\theta(x)$)

$$C_{kk'} = [Cov_\theta(U)]_{kk'} = (\mathrm{E}_\theta(x_k, x_{k'}))_{kk'} \tag{96}$$

Le point que l'on a esquissé à la section précédente, c'est qu'il nous faut identifier la

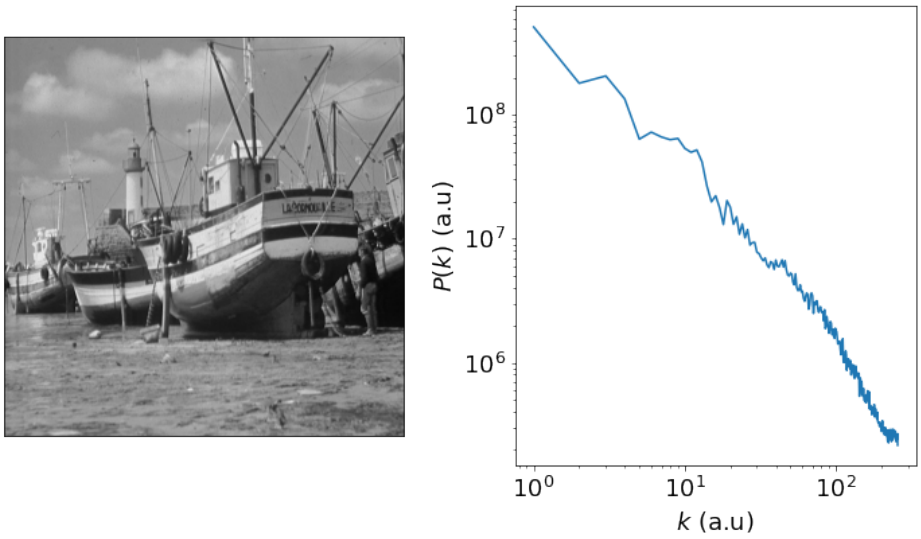

FIGURE 10 – Spectre de puissance d'une image classique, obtenu comme moyenne radiale de la norme au carré de la transformée de Fourier 2D de l'image. Ici k est le nombre d'onde en unité arbitraire. On observe bien une loi en $1/k^2$ à grandes valeurs de k soit les petites échelles de l'image (à petits k on a plutôt ici une dépendance en $1/k$).

base dans laquelle la covariance est diagonale. Si on se place dans *un cas stationnaire*[46] (dans une image cela serait le cas si l'on considère l'invariance par translation) alors

$$\mathrm{E}_\theta(x_k, x_{k'}) = F(k - k') \tag{97}$$

L'invariance par translation (matrice Toeplitz) indique que la base de diagonalisation est celle de Fourier. Une valeur propre σ_k^2 de la matrice de covariance dans cette base s'appelle *la puissance spectrale* ici indicée par ω ($\sigma_\omega^2 = P(\omega)$), et dans le cas d'une image classique le spectre de puissance se comporte en loi de puissance selon $\approx 1/|\omega|^2$ comme sur la figure 10. Donc, l'écart typique entre la plus petite et la plus grande valeur propre est très grand.

La géométrie des réalisations x est celle d'un ellipsoïde en dimension p. Si l'on prend

46. Voir Cours 2021 Sec. 4.4

un petit volume d'iso-probabilité:

$$dV(\alpha) = \{x, \quad 0 \leq \alpha \leq p_\theta(x) \leq (\alpha + d\alpha) \leq 1\} \tag{98}$$

Par exemple, imaginons que l'on soit en dimension 2 et dans la base diagonale $C^{-1} = \text{diag}(\sigma_{min}^{-2}, \sigma_{max}^{-2})$, alors on a bien des couches d'ellipses

$$-2\log(Z(\alpha + d\alpha)) \leq \frac{x_1^2}{\sigma_{min}^2} + \frac{x_2^2}{\sigma_{max}^2} \leq -2\log(Z\alpha) \tag{99}$$

qui sont de plus en plus petites (resp. grandes) quand α tend vers 1 (resp. 0). On voit que ce qui compte, c'est le produit de la valeur de la probabilité et du volume d'iso-proba. Les ellipsoïdes sont *les ensembles typiques* (Eq. 23) introduit par C. Shannon.

4.6 Au delà des champs gaussiens

S. Mallat donne quelques exemples tirés de Mécanique des Fluides (ex. turbulence) et de Cosmologie (ex. gaz interstellaire). Sur la figure 11 nous avons un exemple à gauche d'une image d'un fluide turbulent [47], au centre son spectre de puissance, et à droite un réalisation d'un champ gaussien généré à partir de ce spectre de puissance. Pour se faire, il nous suffit de mesurer la fonction de corrélation à 2 points (transformée de Fourier du spectre de puissance), c'est-à-dire estimer la matrice de covariance. Ce que l'on remarque au premier coup d'œil, c'est que ***le champ gaussien n'a pas de structures comme peut en avoir le champ turbulent***. Et pourtant les modèles gaussiens de turbulence ne sont pas si naïfs que cela. A. Kolmogorov en a établit la base au début des années 1940 [48]. Ce qui est remarquable avec les réseaux de neurones, c'est qu'ils sont capables de reproduire des champs aussi structurés que les vrais. Cependant, le $U(x)$ est beaucoup plus complexe. Néanmoins, comme le dit S. Mallat les physiciens n'ont pas attendu les réseaux de neurones pour aller au-delà de la théorie de Kolmogorov.

Le système qui a été abondamment étudié en Physique Statistique, c'est le modèle d'Ising de réseaux de spins [49]. Sans entrer dans les détails, ce que l'on peut dire c'est que

47. image issue de l'article https://phys.org/news/2015-10-key-features-transition-liquid-smooth.html.
48. Il rédige 4 articles très courts aussi éclairants pour la discipline que ceux de Fisher et Shannon.
49. Le problème que Lars Onsager (1903-76) a résolu exactement en 1944 est le depuis fameux *modèle*

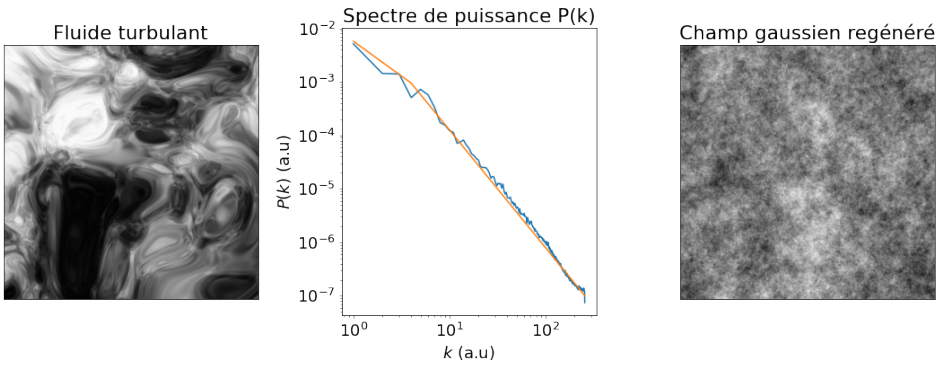

FIGURE 11 – A gauche une image d'un fluide turbulent dans une tuyau (Credit: Piotr Siedlecki/public domain); au centre le spectre de puissance issu de l'image ($\propto k^{-2.2}$); à droite un champ gaussien généré à partir de ce spectre de puissance.

l'interaction de tous les spins sur 1 spin particulier peut être représentée par un potentiel dont la forme peut être modélisée par une forme dite en *"chapeau mexicain"*. Cela donne la théorie dite "$\lambda\phi^4$" qui est aussi utilisée en Physique des Particules pour expliquer la génération des masses de bosons $W^{\pm}$ et Z^0 par le mécanisme de Higgs [50]. Donc, $U(x)$ peut se décomposer de telle façon que

$$\theta \bullet U(x) = \frac{1}{2} x^T C^{-1} x + V(x) \tag{100}$$

avec un terme gaussien et un potentiel $V(x)$ dont la forme est donnée par exemple sur la

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 l'étude des exposants critiques et le développement en Mécanique Statistique de la Théorie des Equations du Groupe de Renormalisation (RGE). Cette théorie 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 de ce que l'on appelle les équations de Callan–Symanzik. Les développements en Mécanique Statistique datent 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 remarquables.

50. On y adjoint d'autres contributeurs à présent et devient le mécanisme de Brout-Englert-Higgs-Hagen-Guralnik-Kibble.

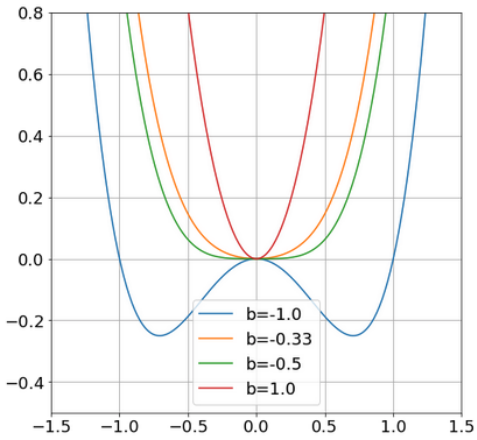

FIGURE 12 – Forme du potentiel $V(x)$ subit par 1 spin dans le cas du modèle d'Ising en ϕ^4 (ou bien il s'agit de l'intensité du pixel d'une image) selon différentes valeurs du paramètre de forme b dont la valeur peut faire apparaitre 2 minima pour $b \leq b_c = -1/2$. Les minima sont localisés en $\pm\sqrt{-(1+2b)/2}$ et la valeur des puits est donnée par $-(1+2b)^2/4$.

figure 12 selon

$$V(x) = x^4 + (1+2b)x^2 \tag{101}$$

Quel est le rôle de ce potentiel $V(x)$? Il est là pour contraindre finalement les valeurs de x à prendre des valeurs comme "piégées" dans les puits de potentiel (négatifs) et augmenter par la même la probabilité $p_\theta(x)$.

Ce modèle d'Ising à permis de comprendre ***les phénomènes de transitions de phases*** qui se manifestent à la limite thermodynamique par une brisure spontanée de symétrie. Pour brosser rapidement le phénomène prenons dans la continuité de ce qui précède une collection de N spins. A *priori* l'énergie du système est invariante par renversement de tous les spins. Par ailleurs, plus la température T du système est grande, plus l'orientation des spins est aléatoire et l'aimantation résiduelle moyenne est nulle. Maintenant, si l'on plonge le système dans un champ extérieur h, ce dernier a tendance à aligner les spins selon une orientation privilégiée: il y a une balance entre cette tendance à l'*ordre* via h, et une tendance au *désordre* via la température. Pour N donné, si on fait tendre h vers 0, on se retrouve dans le cas précédent, il n'y a pas d'aimantation spontanée en moyenne. Mais quand on fait tendre le nombre de spins N vers l'infini (limite thermodynamique),

puis que l'on fait tendre h vers 0, il se trouve que selon la valeur de T (laquelle pourrait gouverner la valeur de b dans le modèle (101)), et surtout si elle devient inférieure à une température critique T_c ($b < b_c$), alors ***l'aimantation spontanée n'est pas nulle, ce qui indique la transition de phase, une brisure de symétrie qui manifeste un effet collectif d'orientation avec des corrélations à longue portée (la longueur de corrélations divergeant à*** $T = Tc$***)***. Autour de $T = T_c$ on peut considérer le système comme instable passant d'une phase à l'autre, avec à chaque transition une discontinuité de symétrie.

Dans les réseaux de neurones, les paramètres θ peuvent être gouvernés par une sorte de température, et l'on peut également assister à des effets collectifs, et ces formes de transitions de phases avec une instabilité du système à leurs abords. Ces changements sont le signe de ***l'instabilité du Hessien*** qui change de comportement avec des ruptures de conditionnement. On a donc des phénomènes qui se retrouve au cœur de l'optimisation et donc touchent beaucoup de domaines.

Maintenant, on peut se demander si l'on peut aller au-delà des modèles de type Ising? La réponse est oui, et c'est là où les réseaux de neurones ont changé la donne. En effet, en utilisant des Generative Models ou des Variational Auto Encodeurs, on peut reproduire des textures complexes telles que des nuages, des tas de cailloux, de bulles, etc [51] (Fig. 13). Le problème est que ces réseaux ont des millions de paramètres et on est "assez loin de comprendre" (sic): pourquoi cela marche? comment relier les paramètres aux interactions de la physique sous-jacente?

Ce sont des problèmes ouverts sur lesquels par exemple S. Mallat et son équipe travaillent, et le point fondamental qui se dégage c'est qu'***il faut comprendre les interactions entre les échelles*** (Voir Cours 2020 Sec. 9.). En découpant l'image en patchs de différentes tailles, avec des petites échelles on se penche sur des interactions hautes fréquences localisées, et à plus grandes échelles on s'informe des interactions basses fréquences de moins en moins localisées. Mais ce qui permet de créer des structures complexes, c'est la façon dont ***les différents niveaux d'échelles interagissent les uns avec les autres***.

Donc, finalement avec le modèle "linéaire" en θ des familles de densité de probabilité, on peut se trouver à représenter des phénomènes très complexes et d'une infinie richesse. Le point crucial là, c'est la modélisation du $U(x)$. Le pendant en Machine Learning, ce

51. Voir Cours 2019 Sec. 2.7. Voir aussi l'article S. Zhang et S. Mallat (2021) https://arxiv.org/pdf/1911.10017.pdf.

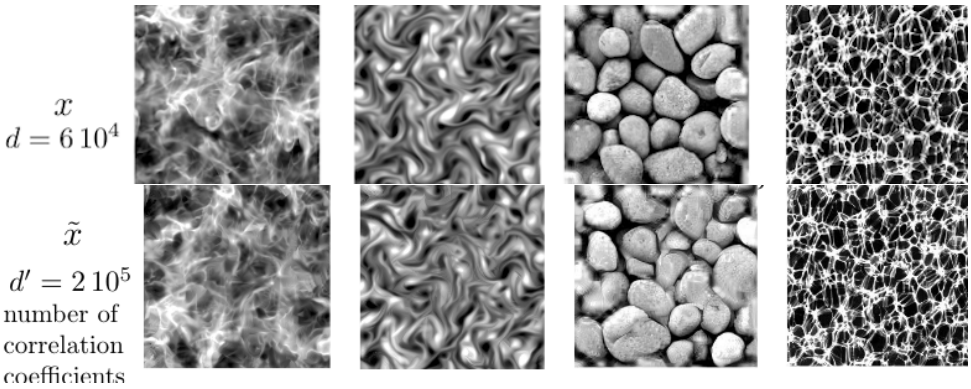

FIGURE 13 – (haut) Exemple de différentes textures: un nuage interstellaire turbulent, un autre type de fluide, un tas de pierres, et des bulles. (bas) Génération de nouvelles textures.

sont ***les modèles à noyaux***[52], pour lesquels le noyau $K(x, x')$ n'est autre que $\{U(x)U^T(x')\}$ soit la matrice de covariance. Une fois que l'on a choisi le noyau, la régression linéaire (*Kernel Ridge Regression*) se fait bien, mais tout ***le problème est d'avoir le (bon) noyau***. Et finalement, on se rend compte des limitations car si cela ne convient pas, que fait-on? Le domaine a été pendant quelque temps bloqué jusqu'à ce que les réseaux de neurones ouvrent une perspective. En effet, on peut les voir comme une façon d'apprendre le bon noyau $U(x)$. Cependant, après s'être rendu compte de l'efficacité des réseaux, on finit par se demander qu'est-ce qu'il y a derrière ces $U(x)$ appris?

4.7 Garantir la consistance

Dans les sections précédentes, on a vu comment obtenir un estimateur optimal, le maximum de vraisemblance, mais on aimerait connaître les conditions qui garantissent la consistance de l'estimateur. C'est-à-dire qu'est-ce qui garantit que lorsque le nombre d'observations tend vers l'infini, alors on converge avec une probabilité 1 vers le bon estimateur qui maximise la vraisemblance en moyenne? Nous complétons ainsi les propriétés de la section 3.5. Examinons les propriétés du maximum de vraisemblance (MLE) définit

52. Voir par ex. Cours 2018 Secs. 7.3, 9.5

par

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \ell(\theta) \tag{102}$$

Théorème 6 *(changement de variable)*
Soit le changement de variable $\eta = g(\theta)$, alors $g(\hat{\theta}) = \hat{\eta}$ est un maximum de vraisemblance (MLE) si $\hat{\theta}$ est un MLE.

Même si la démonstration est plus simple si g est inversible, cela n'est pas requis. Le résultat plus important concerne la consistance.

Théorème 7 *(**consistance du MLE**)*
Soit la vraisemblance

$$\ell(\theta, x) = \log p_\theta(x) \tag{103}$$

avec $x = (x_1, \ldots, x_n)$ iid. On suppose que les observations sont décrites par un certain θ, noté θ^*, qui définit la vraie densité de probabilité. De plus on considère les hypothèses de régularité suivantes:

R0) Si $\theta \neq \theta'$ alors $p_\theta(x) \neq p_{\theta'}(x)$;

R1) Les supports des p_θ sont les mêmes;

R2) θ^* est à l'intérieur de Ω l'ensemble des paramètres.

R2b) de plus on suppose que p_θ est différentiable en θ.

Pour un MLE on a

$$\frac{\partial \ell(\hat{\theta}_n, x)}{\partial \theta} = 0 \tag{104}$$

Cette équation admet potentiellement plusieurs solutions, mais il existe une solution particulière pour laquelle on a la convergence en probabilité, c'est-à-dire

$$\exists \hat{\theta}_n \quad tq. \quad \hat{\theta}_n \xrightarrow[n\to\infty]{prob.} \theta^* \tag{105}$$

C'est un théorème différent de celui que nous avons examiné à la section 3.5 (Th. 3) et nous examinerons la démonstration à la prochaine séance avant d'aborder l'Information de Fisher et les bornes de Cramér-Rao.

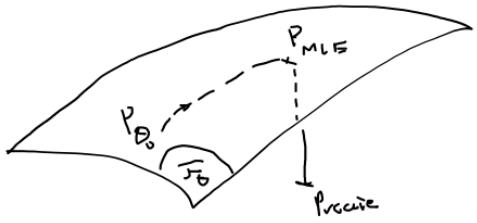

FIGURE 14 – Schématisation de la famille de probabilités $\mathcal{F}_\theta$ et de l'éventuelle erreur si la vraie probabilité sous-jacente aux observations (p_{vraie}) n'appartient pas à cette famille.

5. Séance du 9 Févr.

5.1 Petit préambule

Avant d'aborder la démonstration du théorème 7 faisons un petit préambule. On rappelle que la distribution des données est issue d'une famille paramétrée, c'est-à-dire que $p_{vraie} \in \{p_\theta\}_\theta = \mathcal{F}_\theta$ et on essaye de déterminer le bon θ. Cependant, imaginons que $p_{vraie} \notin \mathcal{F}_\theta$, qu'en est-il si on persiste à utiliser cette famille de distributions? On peut représenter la famille comme une variété, et trouver le MLE se fait en faisant évoluer p_θ sur cette variété. La recherche du MLE est en fait associée à la divergence de Kullback-Leibler [53]:

$$D_{KL}(p||q) := \int p(x) \log \frac{p(x)}{q(x)} dx \tag{106}$$

Elle n'a pas tous les attraits d'une distance, en particulier $D_{KL}(p||q) \neq D_{KL}(q||p)$. Nous verrons dans la 2nd partie du cours que $\log(p)$ est le code optimal pour coder les éléments issus de la loi $p(x)$, et donc la divergence $D_{KL}(p||q)$ mesure une inefficacité de codage que l'on ferait en prenant $\log(q)$ qui lui est optimal pour coder des éléments issus de la loi $q(x)$. Ainsi, si on veut apprécier l'inefficacité à trouver p_{vraie} en prenant les p_θ, cela donne

$$\begin{aligned} D_{KL}(p_v||p_\theta) &= \int p_v(x) \log p_v(x) dx - \int p(x) \log p_\theta(x) dx \\ &= \mathbb{E}_{p_v}[\log p_v] - \mathbb{E}_{p_v}[\log p_\theta] \end{aligned} \tag{107}$$

53. Voir Cours de 2019 Sec. 7.2.3 par exemple.

Donc, ***maximiser la vraisemblance minimise la divergence de Kullback-Leibler***. Mais si $p_{vraie} \notin \mathcal{F}_\theta$ on ne peut atteindre 0, on fait une erreur liée à l'information de projection de p_{vraie} sur $\mathcal{F}_\theta$ (Fig. 14).

5.2 La consistance du MLE

Démonstration 7. On rappelle d'après le théorème 3, pour $x = (x_i)_{i \leq n}$ *iid*,

$$\forall \theta \neq \theta^* \qquad \mathbb{P}\left(\ell(\theta^*, x) > \ell(\theta, x)\right) \xrightarrow[n\to\infty]{prob.} 1 \tag{108}$$

(nb. on a pris ici le log-likelihood). Et on veut montrer qu'il existe une unique séquence de MLE $\hat{\theta}_n$ qui converge en probabilité vers θ^*. Si à chaque étape n il y a plusieurs solutions $\hat{\theta}_n$, on pourra extraire une séquence qui converge vers θ^*. On raisonne en 1D mais cela est généralisable.

Soit $a > 0$ définit tel que $[\theta^* - a, \theta^* + a] \in \Omega$, ce qui est possible d'après l'hypothèse (R2). Soit l'ensemble S_n des observations x définit selon

$$S_n = \{x \ / \ \ell(\theta^*, x) > \max\left(\ell(\theta^* - a, x), \ell(\theta^* + a, x)\right)\} \tag{109}$$

Ce que l'on sait d'après le théorème 3, c'est qu'en probabilité

$$\mathbb{P}(S_n) \xrightarrow[n\to\infty]{prob.} 1 \tag{110}$$

En d'autres termes, presque toutes les observables vont appartenir à S_n.

Dans l'intervalle $[\theta^* - a, \theta^* + a]$, comme $\ell(\theta)$ est dérivable, donc continue, d'après le théorème de Rolle, on trouve une valeur de θ qui annule $\partial_\theta \ell(\theta)$, on la note $\hat{\theta}_n$. Ainsi, définissons l'ensemble $\tilde{S}_n$ des observations

$$\tilde{S}_n = \left\{x \ / \ \exists \hat{\theta}_n, \text{ tq. } \partial_\theta \ell(\hat{\theta}_n, x) = 0 \text{ et } \|\theta^* - \hat{\theta}_n\| < a\right\} \tag{111}$$

Ce que l'on sait c'est que $S_n \subset \tilde{S}_n$ car en effet on ne sait pas a priori pour $x \in \tilde{S}_n$ si $\ell(\hat{\theta}_n, x) < \ell(\theta^*, x)$. Donc $\mathbb{P}(S_n) \leq \mathbb{P}(\tilde{S}_n)$. Ainsi par passage à la limite en probabilité, on a

$$\mathbb{P}(\tilde{S}_n) \xrightarrow[n\to\infty]{prob.} 1 \tag{112}$$

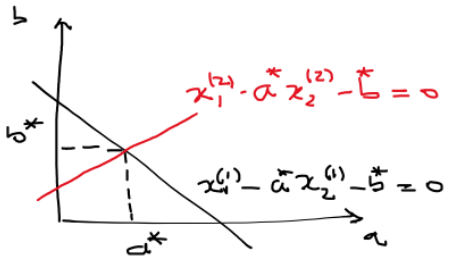

FIGURE 15 – Schématisation de la contrainte de deux observations dans l'espace des paramètres.

Ainsi, pour $\forall a > 0$ en probabilité on trouvera un $\hat{\theta}_n$ proche de θ^*. On a donc notre théorème, à l'ajustement près qu'à chaque étape n, on prend une valeur $\hat{\theta}_n$ s'il y en a plusieurs afin de constituer la séquence qui converge vers θ^*. ■

5.3 Information de Fisher

La question qui vient après avoir montré que le MLE est un estimateur constant, est de savoir si l'on peut faire mieux pour estimer θ^*? On se pose alors la question de ***l'efficacité de l'estimateur MLE***. Pour y répondre on va prendre un estimateur quelconque et obtenir des bornes d'estimations. Cela se fait ici avec la notion d'Information de Fisher.

L'idée sous-jacente est de quantifier la quantité d'information que des observations vont donner sur le paramètre θ.

(NDJE) Voici comment on peut se poser l'idée d'information sur un exemple simple. Soit des observations $(x_1^i, x_2^i)_{i \leq n}$ et imaginons qu'en sous-jacent $x_2^i = a^* x_1^i + b^*$. On se dit qu'en cumulant 2 observations, on a un système à 2 equations à 2 inconnues, et si nos 2 observations sont quelconques, alors Cramer nous donne les valeurs de (a^*, b^*). Mais en fait, posons nous la question dans l'espace des paramètres (a, b) que signifie observer (x_1, x_2)? Il s'agit d'une contrainte ici linéaire comme illustré sur la figure 15. Cette contrainte définissant un lieu géométrique dans l'espace (a, b) est l'information donnée par l'observation. Et la donnée de 2 observations suffit

en effet à déterminer les paramètres du modèle. Si l'on imagine des observables bruitées, alors la contrainte d'une observation n'est pas restreint à une droite mais définit un "tube" comme région de contraintes, et l'intersection de n tubes venant de toutes les observations contraignent alors la détermination des paramètres (a, b) *dans une petite région ellipsoïdale centrée sur* (a^*, b^*).

On va supposer une hypothèse supplémentaire de régularité (R3) qui vient s'ajouter à celles du théorème 7, à savoir que p_θ *est 2 fois dérivable en* θ, et également pour faciliter la démonstration on ajoute (R4) l'hypothèse suivante:

$$\left(\int p_\theta(x)dx\right)'' = \int p''_\theta(x)dx \tag{113}$$

c'est-à-dire que typiquement la dérivée seconde de la probabilité doit être dominée, ce qui est le cas en pratique. Envisageons ***le score***

$$\boxed{s(\theta, x) = \frac{\partial \log p_\theta(x)}{\partial \theta}} \tag{114}$$

qui s'annule qu'en on le calcule pour $\theta = \theta_{MLE}$. Pour $\theta = \theta^*$ qui donne la vraie probabilité de distribution des observables ($p_{vraie} = p_{\theta^*}$), alors

$$\boxed{\mathbb{E}_{x\sim p_{\theta^*}}[s(\theta^*, x)] = 0} \tag{115}$$

En effet,

$$\mathbb{E}_{x\sim p_{\theta^*}}[s(\theta^*, x)] = \int \cancel{p_{\theta^*}(x)} \frac{\partial_\theta p_\theta(x)|_{\theta=\theta^*}}{\cancel{p_{\theta^*}(x)}} dx = \partial_\theta \underbrace{\int p_{\theta^*}(x)dx}_{=1} |_{\theta=\theta^*} = 0 \tag{116}$$

Maintenant que vaut la variance du score? Ceci nous donne une définition de l'Information de Fisher.

Définition 4 *L'**Information de Fisher** est la variance du score*

$$s(\theta, x) = \frac{\partial \log p_\theta(x)}{\partial \theta} \tag{117}$$

calculée en θ^* (nb. le vrai θ). C'est-à-dire

$$\boxed{I(\theta^*) = \mathbb{E}_{x\sim p_{\theta^*}}\left[\left(\frac{\partial \log p_\theta(x)}{\partial \theta}\bigg|_{\theta=\theta^*}\right)^2\right] = Var_{x\sim p_{\theta^*}}[s(\theta^*, x)]} \tag{118}$$

L'idée sous-jacente est que si cette information est importante, alors on est très sensible aux variations de l'estimation du maximum de vraisemblance quand on fait des tirages de données. ***Étant très sensible, on est donc mieux à même de déterminer*** θ^*. Pour exprimer cette intuition, on montre le théorème: suivant

Théorème 8 *(**Information de Fisher et dérivée seconde**)*
L'Information de Fisher est reliée à la courbure du log-likelihood calculée en θ^*, c'est-à-dire

$$\boxed{I(\theta^*) = -\mathbb{E}_{x\sim p_{\theta^*}}\left[\frac{\partial^2 \log p_\theta(x)}{\partial \theta^2}\bigg|_{\theta=\theta^*}\right]} \tag{119}$$

Démonstration 8. La démonstration procède ainsi,

$$\begin{aligned}
\mathbb{E}_{x\sim p_{\theta^*}}\left[\frac{\partial^2 \log p_\theta(x)}{\partial \theta^2}\bigg|_{\theta=\theta^*}\right] &= \int p_{\theta^*} \times \frac{\partial^2 \log p_\theta(x)}{\partial \theta^2}\bigg|_{\theta=\theta^*} dx \\
&= \int \left[-\frac{1}{p_{\theta^*}}(p'_\theta(\theta^*))^2 + p''_\theta(\theta^*)\right] dx \\
&= -\underbrace{\int p_{\theta^*}\left(\frac{\partial \log p_\theta(x)}{\partial \theta}\bigg|_{\theta=\theta^*}\right)^2 dx}_{I(\theta^*)} + \underbrace{\int p''_\theta(\theta^*) dx}_{\left(\int p_\theta(x)dx\right)''_{\theta=\theta^*} = 0}
\end{aligned} \tag{120}$$

■

Regardons à présent l'additivité de l'information de Fisher. Les observations $(x_i)_{i\leq n}$

sont *iid* donc on peut écrire

$$\frac{\partial \log p_\theta(x_1, \ldots, x_n)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log p_\theta(x_i)}{\partial \theta} \tag{121}$$

Les variables $(\partial \log p_\theta(x_i)/\partial\theta)$ sont *iid*, donc la variance s'additionne. Ainsi [54],

$$\boxed{I(\theta^*; x_1, \ldots, x_n) = n\ I(\theta^*; x_i) = n\ I_1(\theta^*)} \tag{122}$$

5.4 Borne de Cramér-Rao

Nous allons utiliser l'information de Fisher pour poser une borne à la précision d'estimation du paramètre θ^*. Si on se place dans un problème d'apprentissage, θ est l'ensemble des paramètres d'un réseau, et l'on veut estimer la précision avec laquelle on les détermine.

Théorème 9 ***(Cramér-Rao)***
Soit $(x_1, \ldots, x_n)$ n observations iid qui sont toutes distribuées[a] *selon $p_\theta(x)$. On considère un estimateur Y de θ qui est une statistique*[b] *des variables $(x_i)_{i \leq n}$*

$$Y = T(x_1, \ldots, x_n) \tag{123}$$

L'espérance de cet estimateur est notée comme suit

$$\mathbb{E}_{x \sim p_\theta}[Y] := \tau(\theta) \tag{124}$$

La variance de Y est alors bornée selon

$$\boxed{Var(Y) \geq \frac{|\tau'(\theta)|^2}{n I_1(\theta)}} \tag{125}$$

où on note $I_1(\theta)$ l'information de Fisher de 1 observation.

54. NDJE: j'ai pris le parti d'ajouter la notation I_1 qui fait référence à 1 observation.

On qualifie ***un estimateur de non biaisé*** *si* $\tau(\theta) = \theta$*, et dans ce cas*

$$Var(Y) \geq \frac{1}{nI_1(\theta)} \tag{126}$$

L'information de Fisher nous donne bien la borne minimale sur la capacité à estimer le paramètre θ *à partir des observations. Notez bien que l'indépendance des observations donne le facteur* $1/n$*.*

a. NDJE: attention, selon le contexte x est soit une observation particulière, soit l'ensemble des observations; de plus ici pour alléger θ est le vrai paramètre.

b. Traditionnellement, le terme "statistique" signifie "fonction".

Démonstration 9. La stratégie de la démonstration passe par le calcul direct de $\tau'(\theta)$. Il vient en utilisant le caractère *iid* des observations et les résultats sur le score (Eqs. 115, 118):

$$\begin{aligned}
\tau(\theta) &= \int p_\theta(x)T(x)dx = \int T(x_1,\dots,x_n)\prod_{i=1}^{n} p_\theta(x_i)\prod_{k=1}^{n} dx_k \\
\Rightarrow \tau'(\theta) &= \int T(x_1,\dots,x_n)\sum_{i=1}^{n}\underbrace{p'_\theta(x_i)\frac{1}{p_\theta(x_i)}}_{(\log p_\theta(x_i))'}\prod_{j=1}^{n} p_\theta(x_j)\prod_{k=1}^{n} dx_k \\
&= \int T(x_1,\dots,x_n)\frac{\partial \log p_\theta(x_1,\dots,x_n)}{\partial\theta}\prod_{j=1}^{n} p_\theta(x_j)\prod_{k=1}^{n} dx_k \\
&= \int T(x)\frac{\partial \log p_\theta(x)}{\partial\theta}p_\theta(x)dx \\
&= \mathbb{E}_{x\sim p_\theta}[Y \times s(\theta,x)] = Cov[Y \times s(\theta,x)] + \mathbb{E}[Y] \times \underbrace{\mathbb{E}[s(\theta,x)]}_{=0}
\end{aligned} \tag{127}$$

Donc, l'inégalité de Cauchy-Schwarz nous dit alors que

$$|\tau'(\theta)|^2 = |Cov_{x\sim p_\theta}[Y \times s(\theta,x)]|^2 \leq Var[Y] \times Var[s(\theta,x)] = Var[Y] \times I(\theta) \tag{128}$$

On a donc le résultat du théorème sachant que $I(\theta) = nI_1(\theta)$. ■

Ce résultat de Cramér-Rao est à la fois très important pour comprendre comment procéder à de l'inférence de paramètres, et assez singulier car il est rare d'avoir une borne

explicite de la limite d'un estimateur. Le travail principal qui mobilise toute l'attention des chercheurs qui font de l'inférence, c'est de ***trouver les estimateurs des paramètres sous-jacent du modèle qui ont la plus grande information de Fisher***. Par exemple, à partir des observations astrophysiques dans tout le spectre électromagnétiques, comment concevoir des observables et les statistiques, pour estimer avec le plus d'efficacité possible (c'est-à-dire la plus grande information de Fisher) les paramètres cosmologiques. Notons que si le modèle standard de base de la Cosmologie (ΛCDM [55]) compte 6 paramètres [56], les inférences portent typiquement sur une centaine de paramètres, comprenant les paramètres de *nuisances* qui portent sur les modélisations d'effets astrophysiques mal connus ainsi que des effets instrumentaux par exemple. Donc, c'est un euphémisme de dire que la tâche n'est pas simple.

Le formalisme précédent se généralise aisément au cas multidimensionnel où $\theta \in \mathbb{R}^d$. On a déjà expérimenté que la dérivée selon θ devient le ***gradient***, ainsi

$$\nabla_\theta \ell(\theta_{MLE}, x) = 0 \qquad\qquad \mathbb{E}_x[\nabla_\theta \ell(\theta^*, x)] = \mathbb{E}_x[s(\theta^*, x)] = 0 \tag{129}$$

et l'information de Fisher devient

$$I(\theta^*) = \mathbb{E}_x[\|\nabla_\theta \ell(\theta^*, x)\|^2) = -\mathbb{E}_x[H[\ell](\theta^*, x)] \tag{130}$$

où apparait le ***Hessien***. Ainsi, ***l'information de Fisher gouverne la vitesse de convergence de l'algorithme de descente de gradient*** (Th. 4) [57]. Par ce biais on voit le lien entre *optimisation* et *précision* à travers la *géométrie* de la surface de la famille des probabilités.

5.5 Optimalité du MLE

Le théorème 9 nous donne une borne, mais la question est de savoir si cette borne peut-être atteinte? Pour cela, nous allons introduire ***l'efficacité d'un estimateur***. Prenons

55. Cold Dark Matter + Constante Cosmologique
56. Planck 2018 https://arxiv.org/abs/1807.06209.
57. NDJE: notons cependant que l'estimation des paramètres ainsi que leurs intervalles de confiances, en dimension *intermédiaire* (ex. astro-cosmo), utilise une autre méthode. Extrait de ref.56: *"[The] nuisance parameters are sampled, along with cosmological parameters, during Markov chain Monte Carlo (MCMC) exploration of the likelihood."*. Ce qui nécessite l'adjonction de *priors*.

le cas d'un *estimateur non-biaisé*, c'est-à-dire tel que $\mathbb{E}(Y) = \theta^*$ [58]. Dans la suite on note l'estimateur $\hat{\theta}_n$ par réminiscence des sections sur le MLE. La première propriété que l'on désirerait, c'est celle de ***consistance*** (Déf. 1), autrement dit une convergence en probabilité de la séquence des $(\hat{\theta}_n)_n$ vers θ^*. Cependant, on veut y ajouter une propriété concernant l'erreur d'estimation: on aimerait quelle atteigne la borne de Cramér-Rao. Considérons alors la variance de $\hat{\theta}_n$ et son rapport avec l'information de Fisher pour définir l'efficacité de l'estimateur:

$$\text{eff.} := \frac{[nI_1(\theta^*)]^{-1}}{Var[\hat{\theta}_n]} \leq 1 \tag{131}$$

La question qui se pose alors est de savoir si l'on peut obtenir une efficacité de 100%? Pour ce faire, nous allons démontrer un résultat sur le MLE qui nous dira que ***la distribution de ce dernier converge vers une loi gaussienne dont la variance est précisément l'information de Fisher***. Il nous faut définir ce que l'on entend par convergence en distribution [59]

Définition 5 *(**Convergence en distribution**)*

Soit une collection de v.a $(x_1, \ldots, x_n)$, la question est de savoir si la loi de ces v.a converge vers la loi d'une variable x? Notons la fonction cumulatrice (ou de répartition) de la probabilité $p_n(x)$ comme

$$F_n(a) = \int_{-\infty}^{a} p_n(x)dx \tag{132}$$

(idem pour F la fonction de répartition de $p(x)$). Ainsi, la convergence en distribution notée

$$p(x_n) \xrightarrow[n\to\infty]{dist.} p(x) \tag{133}$$

signifie que

$$\forall a \; tq. \; F(a) \text{ continue}, \quad \lim_{n\to\infty} F_n(a) = F(a) \tag{134}$$

58. nb. ici on réutilise comme paramètre de la vrai distribution, la notation θ^*.

59. NDJE: en ayant échangé avec S. Mallat, on s'est rendu compte d'une différence dans les notions de convergence en distribution au sens anglo-saxon comme présenté ici, et la convergence en loi française qui stipule que la suite de *v.a* $(X_n)_{n>0}$ dans $\mathbb{R}^d$ converge en loi $X_n \xrightarrow[n\to\infty]{loi} X$ si $\forall f$ continue bornée de $\mathbb{R}^d$ dans $\mathbb{R}$ alors $\mathbb{E}[f(X_n)] \xrightarrow[n\to\infty]{} \mathbb{E}[f(X)]$. Le point est que pour $d = 1$ la convergence en loi est équivalente à la convergence des fonctions de répartition donc à la convergence en distribution, et de plus la convergence en loi est équivalente à la convergence des fonctions génératrices.

Cette définition est en lien avec le théorème Central Limite établit en 1809 par Pierre-Simon de Laplace (1749-1827) généralisant celui de Abraham (de) Moivre (1667-1754) établit sur la distribution de Bernoulli (Jacques -, 1654-1705):

Théorème 10 *(**Central limite**)*
Soit $(x_1, \ldots, x_n)$ *iid avec* $\mathbb{E}(x_i) = \mu$ *et* $0 < Var(x_i) = \sigma^2 < \infty$. *Soit*

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} x_i \tag{135}$$

On sait que (Voir loi faible des grands nombres Th. 1 et démonstration)

$$\bar{X}_n \xrightarrow[n\to\infty]{prob.} \mu, \; Var[\bar{X}_n] = \frac{\sigma^2}{n} \tag{136}$$

Donc, si l'on considère la v.a

$$Z_n := n^{1/2} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \tag{137}$$

sa moyenne est 0 et sa variance est 1, et de plus on a

$$p(Z_n) \xrightarrow[n\to\infty]{dist.} \mathcal{N}(0, 1) \tag{138}$$

S. Mallat indique que la démonstration se fait à partir des fonctions caractéristiques. Elle est laissée aux lecteurs.

Théorème 11 *(**Convergence en loi normale du MLE**)*
On reprend les hypothèses de régularités du théorème de consistance du MLE (7), ainsi que celles sur la dérivée seconde (Eq. 113) auxquelles on ajoute une nouvelle hypothèse (R5)

$$|(\log p_\theta(x))'''_\theta| < M(x) \tag{139}$$

telle que $\mathbb{E}[M(x)] < \infty$ *ce qui va permettre d'appliquer le théorème de convergence dominée aux termes d'erreur. Etant donné ces différentes hypothèses, on a que pour toute séquence de MLE* $\hat{\theta}_n$ *qui converge en probabilité vers* θ^* *(on sait qu'il y en a*

une au moins), alors

$$\boxed{p(\sqrt{n}(\hat{\theta}_n - \theta^*)) \xrightarrow[n\to\infty]{dist.} \mathcal{N}(0, I^{-1}(\theta^*) = [nI_1(\theta^*)]^{-1})} \tag{140}$$

(nb. en dimension n, on prend la matrice inverse du Hessien). Ce théorème nous dit que ***le MLE est un estimateur asymptotiquement optimal*** *puisqu'il atteint la borne de Cramér-Rao.*

Démonstration 11. Nous allons esquisser les étapes de la démonstration. Si l'on prend un MLE $\hat{\theta}_n$, on sait que la dérivée du log-likelihood y est nulle, donc en effectuant un DL au voisinage de $\hat{\theta}_n$ calculé en θ^*; il vient

$$\ell'(\theta^*) = \cancel{\ell'(\hat{\theta}_n)} + (\theta^* - \hat{\theta}_n)\ell''(\hat{\theta}_n) + \dots \tag{141}$$

En $\ell'(\theta^*)$ on reconnait le score de n *v.a iid* dont l'espérance est nulle (Eq. 115) et de variance égale à l'Information de Fisher (Eq. 118) ($I(\theta^*) = nI_1(\theta^*)$). Par le théorème centrale limite on obtient

$$n^{1/2}\left(\frac{\ell'(\theta^*)}{I^{1/2}(\theta^*)}\right) \xrightarrow[n\to\infty]{dist.} \mathcal{N}(0,1) \Rightarrow n^{1/2}\ell'(\theta^*) \xrightarrow[n\to\infty]{dist.} \mathcal{N}(0, I(\theta^*)) \tag{142}$$

Donc, il nous faut considérer à présent

$$n^{1/2}(\hat{\theta}_n - \theta^*) = -\frac{n^{1/2}\ell'(\theta^*)}{\ell''(\hat{\theta}_n)} \tag{143}$$

et traiter le dénominateur, lequel est le plus délicat. Attention en effet l'information de Fisher peut s'écrire avec les dérivées secondes (Th. 8) ($-\ell''(\theta^*) = I(\theta^*)$) mais calculées au point θ^* et non $\hat{\theta}_n$. Si on fait un DL au voisinage de θ^*, on obtient

$$\ell''(\hat{\theta}_n) = \ell''(\theta^*) + (\hat{\theta}_n - \theta^*)\ell'''(\theta^*) + \dots \tag{144}$$

La condition que le terme en $\ell'''(\theta^*)$ va tendre vers 0 tient en l'hypothèse R5. Imaginons pour un temps qu'on néglige ce terme alors [60]

$$n^{1/2}(\hat{\theta}_n - \theta^*) \underset{n\to\infty}{\approx} \frac{n^{1/2}\ell'(\theta^*)}{I(\theta^*)} \xrightarrow[n\to\infty]{dist.} \mathcal{N}(0, I^{-1}(\theta^*) = [nI_1(\theta^*)]^{-1}) \tag{145}$$

Ce qui nous donne le théorème. Donc le point technique est le contrôle des termes en $\ell'''(\theta^*)$. Le théorème se généralise en dimension d grâce à l'extension de la convergence dominée. ■

Une fois que l'on a ce résultat, on se rend compte que l'on a accès aux ***intervalles de confiance*** (contours en 2D). C'est à la fois très important en Physique où le travail essentiel consiste non seulement à produire des résultats sur tel ou tel paramètre, mais également à donner la probabilité de trouver le dit paramètre dans les bornes à X% niveau de confiance. Qui plus est la communauté s'émeut dès lors qu'il y a des *tensions* à n-sigma entre plusieurs expériences donnant des résultats sur les mêmes paramètres. Mais, cette problématique d'estimation des intervalles de confiance tend à venir en apprentissage. Par exemple si on fait de la régression logistique, l'on sait qu'il y a unicité de θ^* (convexité), et donc on sait qu'en théorie quand n tend vers l'infini, $\hat{\theta}_n$ est consistant (convergence en probabilité) et on a un intervalle de confiance grâce à la convergence en distribution. Malheureusement, dans le cas des réseaux de neurones cette technique ne marche pas pour pas mal de raisons. En voici deux:

- il n'y a pas du tout d'unicité de θ^* car on n'a ***pas de convexité***;
- le point est que le formalisme implicitement fait l'hypothèse que d est fixé et $n \gg d$ (***régime de statistique classique***), or le nombre (colossal) de paramètres dépasse largement le nombre d'échantillons, et cette l'***over-paramétrisation*** $d \gtrsim n$ est très effective (Voir Sec. 2.1). L'exemple que donne S. Mallat concerne l'évaluation d'une PCA sur des images où typiquement on a $d = k^2$ le nombre de pixels $O(10^6)$. Or, le nombre d'échantillons auxquels on a accès est sans doute bien moindre. Donc, ici on ne peut sans doute pas espérer estimer de manière "consistante" toute la base PCA. Dans ce cas qu'est ce qui reste de consistant dans cette estimation partielle? Dans les meilleurs des cas on aura accès aux plus grandes valeurs propres (et vecteurs associés). Ainsi, il va nous falloir sortir du cadre classique en statistiques, ce qui

60. si $u \sim \mathcal{N}(\mu, \sigma^2)$ alors $u/a \sim \mathcal{N}(\mu, \sigma^2/a^2)$.

donne un champ nouveau d'exploration en ML.

Dans la seconde partie du cours, nous allons emprunter une vision différente, en suivant les pas de C. Shannon. Cela consistera à donner une information ***indépendante des paramètres***. Le problème de C. Shannon n'était pas de découvrir les paramètres de tel ou tel phénomène physique, mais de transmettre des données le plus efficacement possible. Donc, il se posait la question de savoir quel est le nombre minimal de bits nécessaires pour transmettre une information? Même si *a priori* on est loin de l'information de Fisher, on va se rendre compte qu'il y a une convergence entre ces notions, et cela fera ressortir un concept bien connu en Physique Statistique à savoir l'***entropie***.

6. Séance du 16 Févr.

6.1 Introduction

Nous allons aborder le point de vue de Cl. Shannon sur la notion d'information (1948). Comme déjà dit (Sec. 2.5), il s'agit de trouver ***l'information intrinsèque*** contenue dans les observations, sans faire référence à un quelconque modèle sous-jacent. Cette information intrinsèque est reliée au ***nombre minimum de bits*** qu'il est nécessaire d'utiliser pour la coder ou la transmettre sur des canaux, processus qui peut engendrer des erreurs. Ce qui est remarquable, c'est que le travail de Shannon a ouvert des ponts vers la Physique Statistique via la notion d'***Entropie*** qui spécifie le nombre de configurations d'un système. Dans les années 60, A. Kolmogorov va de nouveau se poser la question de Shannon selon le point de vue du nombre minimum d'information qu'il faut pour reproduire des observations. Cependant, A. Kolmogorov utilise une ***Machine de Turing*** en définissant la complexité (ou information de Kolmogorov) comme la taille du programme minimum qui permet de reproduire une séquence. Il y a une correspondance entre les deux notions en considérant les processus ergodiques stationnaires [61], car alors il y a équivalence entre l'entropie de Shannon et la quantité d'information de Kolmogorov (à une constante prés).

61. (NDJE) Petit point de vocabulaire: 1) un processus est *ergodique* si ses propriétés statistiques peuvent être étudiées à partir d'*une seule réalisation* suffisamment longue (ex. ergodicité sur la moyenne où moyennes temporelles tendent vers les moyennes d'ensemble); 2) un processus est *stationnaire* si ses propriétés statistiques caractérisées par des espérances mathématiques sont *indépendantes du temps*. Ces deux notions ne sont pas identiques. Si $X(t) = x_0 + n$ avec x_0 une constante et n une *v.a* alors $\mathbb{E}[x(t)] =$

Pourquoi s'intéresser à l'information de Shannon plutôt qu'à celle de Kolmogorov? On peut le justifie par le fait que l'information de Kolmogorov à part quelques cas est très difficile à calculer, alors que l'entropie de Shannon non seulement elle est intuitive mais aussi on peut l'estimer à partir des observations. Nous allons donc aborder cette notion d'entropie, montrer *son additivité* qui permet de la relier à la notion d'information, et nous verrons ces ***phénomènes de concentration*** qui sont à la base de la théorie de Shannon. En effet, selon lui la raison pour laquelle l'entropie permet effectivement de quantifier la taille minimale d'un code qui reproduirait les observations, c'est que géométriquement ces observations se concentrent dans des ***ensembles typiques*** (Eq. 23, Fig. 5) dont ***la taille est spécifiée par l'entropie***. Donc, en comptant le nombre d'éléments de ces ensembles, on a accès au nombre de bits nécessaires pour les coder. In fine, à partir de l'entropie on pourra définir des modèles, ce qui bouclera sur l'information de Fisher, par le biais de ***modèles d'entropie maximum***, un principe élaboré en 1957 par Edwin Thompson Jaynes (1922-98), où l'on établira un pont avec le maximum de vraisemblance. Une des applications de cette théorie du codage est ***la compression des signaux*** et les notions de ***distorsion*** versus compression. Par ce biais en sous-jacent on se pose la question où sont les structures du signal et de leur ***représentation*** pour appréhender la géométrie des ensembles typiques.

6.2 L'entropie de Shannon

On se place dans le cadre où on a un *alphabet* de taille fini, noté $\mathcal{A} = \{a_k\}_{1 \leq k \leq K}$, où les symboles a_k sont les valeurs que prend une variable aléatoire X avec une probabilité $p(a_k)$ pour chacun d'entre-eux. Peut-on connaître l'incertitude sur la valeur de x une réalisation de X? Imaginons que les probabilités $p(a_k)$ soient toutes identiques (eg. $1/K$) alors on a en quelque sorte une *incertitude maximale* sur la valeur de x, c'est-à-dire aucun symbole particulier est privilégié. A l'inverse si $p(a_{k_0} = 1)$ alors on connait avec certitude quel est le résultat de x. D'une certaine mesure, la variance nous donnerait accès à l'erreur sur x, mais elle est surtout reliée à ***la notion de codage***. En effet, imaginons une distribution concentrée sur peu de symboles, il est tentant de vouloir exprimer avec ***peu de bits ces symboles privilégiés*** qui sont souvent utilisés, tout en se permettant l'usage

$x_0 + \mathbb{E}[n]$ ne dépend pas du temps, c'est donc un processus stationnaire, par contre il n'est pas ergodique car par ex. soit $x_i(t)$ une réalisation qui fixe la valeur de n à la valeur n_i, alors $\frac{1}{2T}\int_{-T}^{T} x_i(t)dt \to x_0 + n_i$ ce résultat dépend de la réalisation donc le processus n'est pas ergodique.

d'un ***maximum de bits pour les symboles peu utilisés***. Ceci est particulièrement effectif quand on veut coder des phrases du langage naturel où les "symboles" sont les mots d'un corpus de vocabulaire.

Définition 6 *(**Entropie de Shannon**)*
L'entropie au sens de Shannon est donnée par

$$\boxed{\mathbb{H}(X) := -\mathbb{E}_{x\sim p}[\log p(X)] \geq 0} \tag{146}$$

qui dans le cas d'une v.a X à valeurs dans un alphabet $\mathcal{A} = \{a_k\}_{1\leq k\leq K}$ donne[a]

$$\mathbb{H}(X) = -\sum_{k=1}^{K} p(X = a_k) \log p(X = a_k) \tag{147}$$

a. on utilise le "log" sans préciser la base.

NDJE: Au passage, je note l'entropie avec la lettre $\mathbb{H}$ 1) car Cl. Shannon a utilisé la lettre H (rappel: Boltzmann utilisait la lettre S) et 2) il fallait la différentier du "H" de Hessien. Si nous reprenons les deux exemples extrêmes mentionnés ci-dessus, dans le cas d'une distribution uniforme alors $\mathbb{H}[\mathcal{U}] = \log K = \log |\mathcal{A}|$ et dans le cas d'une distribution piquée sur 1 seul symbole alors $\mathbb{H}[\delta] = 0$. Nous verrons que $\log |\mathcal{A}|$ est la borne supérieure de $\mathbb{H}(X)$ (Eq. 4). Ainsi on comprend que l'entropie de Shannon mesure bien une erreur sur la valeur d'une réalisation x.

Considérant deux *v.a* , on définit l'entropie jointe et l'entropie conditionnelle comme suit

Définition 7 *L'entropie jointe de deux v.a X, Y à valeurs dans $\mathcal{A}$*

$$\mathbb{H}(X, Y) := -\mathbb{E}_{(x,y)\sim p}[\log p(X, Y)] = -\sum_{k,k'} p(X = a_k, Y = a_{k'}) \log p(X = a_k, Y = a_{k'}) \tag{148}$$

Définition 8 *L'entropie conditionnelle de deux v.a X, Y à valeurs dans $\mathcal{A}$*

$$\begin{aligned}\mathbb{H}(Y|X) &:= -\sum_k p(X=a_k)\mathbb{H}(Y|X=a_k)\\ &= -\sum_{k,k'} p(X=a_k, Y=a_{k'})\log p(Y=a_{k'}|X=a_k)\\ &= -\mathbb{E}_{(x,y)\sim p}[\log(Y|X)] \end{aligned} \tag{149}$$

Par la suite on pourra simplifier la notation soit par exemple en utilisant $p(a_k) = p(X = a_k)$, et $p(a_k, a_{k'}) = p(X = a_k, Y = a_{k'})$, soit en faisant référence à $p(x, y)$, $p(x)$, $p(y|x)$ qui prête moins à confusion. Les deux entropies précédentes sont reliées selon

Propriété 2

$$\boxed{\mathbb{H}(X,Y) = \mathbb{H}(X) + \mathbb{H}(Y|X)} \tag{150}$$

En effet,

$$\begin{aligned}\mathbb{H}(X,Y) - \mathbb{H}(X) &= -\sum_{x,y} p(x,y)\log p(x,y) + \sum_x p(x)\log p(x)\\ &= -\sum_{x,y} p(x,y)\log p(x,y) + \sum_x \left(\sum_y p(x,y)\right)\log p(x)\\ &= -\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)} = -\sum_{x,y} p(x,y)\log p(y|x)\\ &= \mathbb{H}(Y|X)\end{aligned} \tag{151}$$

Au passage on se dit que si l'entropie mesure une incertitude, la relation ci-dessus est tout à fait intuitive.

Voyons de ce qu'il en est de l'additivité de l'entropie ainsi définie. Pour cela on va se servir de la notion d'information mutuelle par le biais de la divergence de Kullback-Leibler (Voir note 53) ou ***entropie relative***. C'est un outil très utile en probabilité.

6.3 Entropie relative et Information mutuelle

Rappelons la définition de la divergence de Kullback-Leibler:

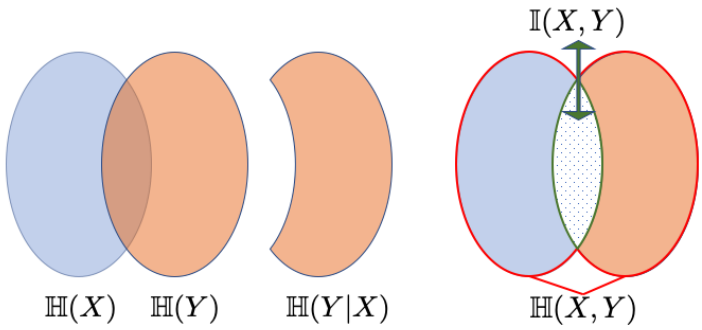

FIGURE 16 – Schématisation d'entropie $\mathbb{H}(X)$ et $\mathbb{H}(Y)$, de $\mathbb{H}(Y|X)$, ainsi que de l'information mutuelle (Eq. 153) et de l'entropie jointe (Eq. 150).

Définition 9 *(**Kullback-Leibler**)*
Si le support[a] *de q inclut le support de p alors*

$$D(p\|q) := \sum_x p(x) \log \frac{p(x)}{q(x)} < \infty \tag{152}$$

La somme pouvant se transformer en intégrale si besoin.

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

Quant à l'information mutuelle qui va nous donner une ***mesure d'indépendance*** définissons là à partir de la divergence de Kullback-Leibler comme suit.

Définition 10 *(**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

$$\mathbb{I}(X,Y) := D(p(x,y)\|p(x)p(y)) \tag{153}$$

Le lien avec l'entropie s'exprime selon la propriété suivante

Propriété 3

$$\begin{aligned}\mathbb{I}(X,Y) &= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \\ &= \mathbb{H}(X) - \mathbb{H}(X|Y) \\ &= \mathbb{H}(Y) - \mathbb{H}(Y|X)\end{aligned} \tag{154}$$

En effet, il vient simplement

$$\begin{aligned}\mathbb{I}(X,Y) + \mathbb{H}(X,Y) &= -\sum_{x,y} p(x,y)\log[p(x)p(y)] \\ &= -\sum_x \underbrace{\sum_y p(x,y)}_{p(x)} \log p(x) - \sum_y \underbrace{\sum_x p(x,y)}_{p(y)} \log p(y) \\ &= \mathbb{H}(X) + \mathbb{H}(Y)\end{aligned} \tag{155}$$

ensuite on utilise Eq. 150. En quelque sorte ***l'information mutuelle est mesurée par l'impact que la connaissance de*** x ***donne sur la réduction d'incertitude sur la valeur*** y (et vice versa en changeant le rôle de x et y). Si les deux variables sont indépendantes la réduction d'incertitude est nulle. Ces différentes notions peuvent se schématiser comme sur le figure 16.

Pour démontrer un certain nombre de résultats, on a besoin de l'inégalité de Jensen dans le contexte des probabilités [62]

Théorème 12 *(**Inégalité de Jensen**)*
*Soit f une **fonction convexe** en dimension 1 (dérivée seconde positive ou nulle), alors pour toute v.a X*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]) \tag{156}$$

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

Nous allons nous en servir pour montrer que la divergence de Kullback-Leibler est positive.

62. nb. voir aussi Eq. 48.

Théorème 13 *(**Positivité de Kullback-Leibler**)*

$$\boxed{D(p\|q) \geq 0} \qquad \boxed{D(p\|q) = 0 \quad ssi \quad p(x) = q(x) \quad \forall x} \tag{157}$$

Démonstration 13. La fonction log est strictement concave donc $-\log$ strictement convexe

$$\begin{aligned} D(p\|q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} = -\sum_x p(x) \log \frac{q(x)}{p(x)} \\ &= \mathbb{E}_p \left[-\log \frac{q(x)}{p(x)} \right] \geq -\log \mathbb{E}_p \left[\frac{q(x)}{p(x)} \right] = -\log(1) = 0 \end{aligned} \tag{158}$$

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 $\sum_x p(x) = \sum_x q(x) = 1$ alors $c = 1$, et donc on a le second résultat du théorème. ■

Donc, la divergence de Kullback-Leibler donne bien une sorte de "distance" entre les deux probabilités p et q au sens où elle donne un signal de similarité si elle est proche de 0. Cependant, ce n'est pas une distance car $D(p\|q) \neq D(q\|p)$.

Une conséquence de la positivité de $D(p\|q)$ concerne l'information mutuelle, puisqu'elle lui est directement reliée par définition. Ainsi

$$\boxed{\mathbb{I}(X, Y) \geq 0} \tag{159}$$

et on a une égalité *ssi* en cas d'indépendance:

$$\boxed{\mathbb{I}(X, Y) = 0 \quad \text{ssi} \quad X, Y \text{ indépendantes}} \tag{160}$$

Une conséquence, concerne l'entropie d'une variable aléatoire qui prend ses valeurs dans un alphabet (voir les deux exemples Sec. 6.2).

Propriété 4 *Soit X une v.a à valeurs dans $\mathcal{A}$ de taille fini alors*

$$\boxed{\mathbb{H}(X) \leq \log |\mathcal{A}|} \tag{161}$$

En effet, soit $\mathcal{A} = \{a_k\}_{1 \leq k \leq K}$, et soit la loi uniforme sur cet alphabet, $q(a_k) = 1/K \ \forall k$, alors pour n'importe quelle *v.a* X de probabilité $p(x)$

$$0 \leq D(p\|q) = \sum_k p(a_k) \log p(a_k) + \sum_k p(a_k) \log K = -\mathbb{H}(X) + \log K \tag{162}$$

Une autre propriété intuitive si l'on comprend l'entropie comme une erreur, est que si on ajoute de l'information en conditionnant une *v.a* alors

Propriété 5

$$\boxed{\mathbb{H}(X|Y) \leq \mathbb{H}(X)} \tag{163}$$

Elle est évidente, quand on se souvient des relations entre l'information mutuelle et l'entropie (Eqs. 3) et le fait que l'information mutuelle est toujours positive. Une manière de se représenter cette relation est donnée sur la figure 16 où $\mathbb{H}(Y|X)$ est le crossant orange de taille plus petite que $\mathbb{H}(Y)$.

Donc finalement, l'entropie de Shannon répond à notre intuition sur ce que devrait être une erreur sur un processus aléatoire. Le point très important que nous ne démontrons pas ici, c'est que réciproquement, si on se donne les relations ci-dessus et que l'on se pose quelle doit être la forme de $\mathbb{H}(X)$ alors on tombe sur l'entropie de Shannon.

(NDJE) Dans son article de 1948, Shannon donne 3 conditions pour la fonction $\mathbb{H}(p_1, p_2, \ldots, p_K)$ où $(p_k)_k$ sont les probabilités connues de K événements:

1) $\mathbb{H}$ doit être une fonction continue de toutes ses variables p_k;

2) dans le cas d'équiprobabilité $p_k = 1/K$ $(\forall k)$ alors $\mathbb{H}$ doit être une fonction monotone de K, reflétant le fait que lorsqu'il y a plus de choix il y a en contrepartie plus d'incertitude;

3) Dans le cas où le problème original est subdivisé en sous-problèmes, alors la fonction $\mathbb{H}$ originale doit être une somme pondérée des fonctions $\mathbb{H}$ des sous-problèmes.

Sous ces conditions il démontre que

$$\mathbb{H} = -C \sum_k p_k \log p_k \tag{164}$$

avec C une constante positive qu'il prend égale à 1, ce qui est une sorte de choix d'unité (d'ailleurs on ne précise pas quelle type de log est envisagé).

L'entropie de Shannon répond donc à nos attentes en termes d'incertitude sur des processus, mais où elle devient très puissante c'est quand on fait le lien avec ***les phénomènes de concentration***.

6.4 Ensembles typiques

Quand on veut faire de la modélisation d'observations, on essaye de se représenter ***la géométrie de l'espace*** dans lesquelles elles évoluent. Et comme nous l'avons déjà évoqué (Sec. 2.2), en grande dimension où le point de vue probabiliste est le plus souvent plus puissant que le point de vue déterministe, c'est bien à cause des phénomènes de concentration (Fig. 3). En ML souvent on parle de "variétés" mais ne nous y trompons pas, ces ensembles ne sont pas forcément différentiables. On aimerait donc caractériser la géométrie de ces ensembles, en calculer la taille que l'on pense être beaucoup plus petite que la taille de l'ensemble dans lequel ils baignent. Et dans ce contexte, nous allons voir que ***l'entropie caractérise le volume des ensembles typiques*** (ou bien le nombre d'éléments dans le cas discret) (Eq. 23, Fig. 5).

Plaçons nous dans un cas où on a des réalisations *iid* $(x_1, x_2, \ldots, x_n)$ d'un processus X. La probabilité de ces réalisations est bien entendu

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i) \tag{165}$$

et comme dans le cas de la notion de vraisemblance, on a envie de prendre le logarithme de cette expression. Si l'on pondère par $1/n$, on a

$$\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \log p(x_i) \tag{166}$$

c'est-à-dire que l'on a une moyenne de *v.a iid*. Donc, d'après le Th. 1 de la convergence en probabilité, il vient

$$\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) \xrightarrow[n\to\infty]{prob.} \mathbb{E}_{x\sim p(x)}[\log p(x)] = -\mathbb{H}(x) \tag{167}$$

Donc [63]

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} \mathbb{P}\left(\left|-\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) - \mathbb{H}(x)\right| \leq \varepsilon\right) = 1 \tag{168}$$

Ainsi, 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, $\{x_i\}_{1\leq i\leq n} = \{x\}$

$$\boxed{T_n^\varepsilon = \left\{\{x\} \in \mathcal{A}^n, \left|-\frac{1}{n}\log p(\{x\}) - \mathbb{H}(x)\right| \leq \varepsilon\right\}} \tag{169}$$

Et donc pour tout $\varepsilon > 0$, et n assez grand

$$\mathbb{P}[\{x\} \in T_n^\varepsilon] \geq 1 - \varepsilon \tag{170}$$

c'est-à-dire que ***presque toutes les réalisations vont appartenir à*** T_n^ε***, d'où l'appellation d'ensemble typique***. La question qui vient est: quelle est leur taille?

Si on réécrit la contrainte définissant T_n^ε et utilisant ***le log à base*** 2 (ce qui fixe la constante c mentionnée ci-dessus), alors $\forall \varepsilon > 0$

$$\left|-\frac{1}{n}\log p(\{x\}) - \mathbb{H}(x)\right| \leq \varepsilon \Rightarrow \boxed{2^{-n(\mathbb{H}(x)+\varepsilon)} \leq p(\{x\}) \leq 2^{-n(\mathbb{H}(x)-\varepsilon)}} \tag{171}$$

Notons que par additivité de l'entropie $n\mathbb{H}(x) = \mathbb{H}(\{x\})$ qui est ***une constante indépendante d'une réalisation*** $\{x\}$ ***quelconque***. Donc, ce qui remarquable c'est que à ε-près l***a probabilité est quasi constante*** sur ces ensembles typiques. Ainsi, ***les observations se concentrent dans ces ensembles, et en même tant se retrouvent n'importe où à l'intérieur*** ce qui est une conséquence de l'indépendance. On appelle cela, ***l'équipartition asymptotique*** [64]. On peut donc énoncer les deux propriétés suivantes des ensembles ty-

63. NDJE: Pour la définition des ensembles typiques Eq. 23, la notation était légèrement différente: $(x_1, \ldots, x_n) = \{x\}$, et $\mathbb{H}[p]$ est en fait $\mathbb{H}(x)$. Ici j'ai pris une notation en lien avec la définition de l'entropie des sections de cette séance.

64. NDJE: ce qui est à mettre en regard vis-à-vis du ***principe fondamental d'équiprobabilité*** des micro-

piques:

Propriété 6 *Outre le fait que*

$$\boxed{\mathbb{P}[\{x\} \in T_n^\varepsilon] \geq 1 - \varepsilon} \tag{172}$$

le cardinal de l'ensemble T_n^ε *satisfait*

$$\boxed{(1-\varepsilon)2^{n(\mathbb{H}(x)-\varepsilon)} \leq |T_n^\varepsilon| \leq 2^{n(\mathbb{H}(x)+\varepsilon)}} \tag{173}$$

ce qui conduit à dire que le nombre d'éléments $\{x\}$ *de l'ensemble est à* ε*-prés* $2^{n\mathbb{H}(x)}$ *(soit aussi* $2^{\mathbb{H}(\{x\})}$*).*

La seconde propriété se démontre ainsi. Soit les $\{x\} \in \mathcal{A}^n$, en utilisant la relation 171 on a

$$1 = \sum_{\{x\}\in\mathcal{A}^n} p(\{x\}) \geq \sum_{\{x\}\in T_n^\varepsilon} p(\{x\}) \geq \sum_{\{x\}\in T_n^\varepsilon} 2^{-n(\mathbb{H}(x)+\varepsilon)} = |T_n^\varepsilon| \times 2^{-n(\mathbb{H}(x)+\varepsilon)} \tag{174}$$

ce qui donne une des deux inégalités. Considérant la première propriété, on a

$$\sum_{\{x\}\in T_n^\varepsilon} p(\{x\}) \geq 1 - \varepsilon \tag{175}$$

donc en utilisant Eq. 171, il vient

$$1 - \varepsilon \leq \sum_{\{x\}\in T_n^\varepsilon} 2^{-n(\mathbb{H}(x)-\varepsilon)} = |T_n^\varepsilon| \times 2^{-n(\mathbb{H}(x)-\varepsilon)} \tag{176}$$

ce qui donne l'autre inégalité. A partir de ces propriétés nous allons pouvoir faire du codage.

états d'un système thermodynamique d'énergie dans l'intervalle $[E, E + dE]$.

6.5 Code typique

Pourquoi peut-on faire du codage? Imaginons que j'ai[65] $X = (x_1, x_2, \ldots, x_n)$ qui a n coordonnées sachant que ces coordonnées sont des *v.a iid* régies par une loi $p(x)$ où les valeurs de x son prises soit dans un alphabet de taille finie K, ou bien sur un intervalle de $\mathbb{R}$, en tous les cas pour laquelle je peux calculer $\mathbb{H}(x)$. On associe à X un mot binaire $w(X)$ lequel a une certaine longueur $\ell(X)$ (ou $\ell(w(X))$), et l'on regarde la longueur moyenne (ou nombre de bits) par symbole

$$R = \frac{1}{n} \sum_X \ell(X) p(X) \tag{177}$$

Le mot $w(X)$ peut varier d'une observation X à l'autre et donc on veut connaitre la taille minimale en moyenne.

Maintenant, connaissant l'existence de l'ensemble typique associé à $p(x)$, on se dit que soit $X \notin T_n^\varepsilon$ mais cela va être le cas avec une très faible probabilité ($\leq \varepsilon$), soit $X \in T_n^\varepsilon$ avec une forte probabilité laquelle est quasi-uniforme sur l'ensemble. L'idée est donc de prendre des codes plus courts quand la probabilité est grande et plus long quand la probabilité est petite (rappelons nous du codage du langage naturel, Sec. 6.2). Or, pour $X \in T_n^\varepsilon$, les éléments étant équiprobables, il est naturel de prendre un code de même longueur pour ces éléments. La taille doit permettre de distinguer tous les éléments de cet ensemble typique, elle est donc proche de $\log_2 |T_n^\varepsilon|$. Ainsi, définissons ***le code typique ou ε-typique***:

Définition 11 *(**code ε-typique**)*

- *Si $X \in T_n^\varepsilon$, $\ell(X) = \lceil n(\mathbb{H}(x) + \varepsilon) \rceil = \lfloor n(\mathbb{H}(x) + \varepsilon) \rfloor + 1$*
- *Si $X \notin T_n^\varepsilon$, en se rappelant la taille de cet ensemble, $\ell(X) = \lfloor n \log_2 K \rfloor + 1$*
- *il nous ajouter 1 bit à chaque $\ell(X)$ pour signaler que X est dans l'ensemble typique ou non.*

On peut alors borner R, selon

65. NDJE: ici X représente ce qu'auparavant était noté $\{x\}$.

Théorème 14 ***(Borne de Shannon)***
$\exists C$ tq. $\forall \varepsilon > 0$, pour n assez grand et un codage ε-typique alors le nombre moyen de bits par symbole satisfait

$$\boxed{R \leq \mathbb{H}(x) + C\varepsilon} \tag{178}$$

Démonstration 14. Ecrivons l'expression de R selon

$$\begin{aligned} R &= \frac{1}{n} \sum_{X \in T_n^\varepsilon} \ell(X) p(X) + \frac{1}{n} \sum_{X \notin T_n^\varepsilon} \ell(X) p(X) \\ &= \frac{1}{n} (\lfloor n(\mathbb{H}(x) + \varepsilon) \rfloor + 2) \left(\sum_{X \in T_n^\varepsilon} p(X) \right) + \frac{1}{n} (\lfloor n \log_2 K \rfloor + 2) \left(\sum_{X \notin T_n^\varepsilon} p(X) \right) \end{aligned} \tag{179}$$

Or $\sum_{X \in T_n^\varepsilon} p(X) \leq 1$ et $\sum_{X \notin T_n^\varepsilon} p(X) \leq \varepsilon$, comme $\lfloor x \rfloor \leq x$ donc

$$R \leq \frac{1}{n}(n(\mathbb{H}(x) + \varepsilon) + 2) + \frac{1}{n}(n \log_2 K + 2)\varepsilon \leq \mathbb{H}(x) + \varepsilon \left(\frac{3}{n} + \log_2 K \right) + \frac{2}{n} \tag{180}$$

ce qui permet la borne supérieure et d'identifier C en justifiant le n suffisamment grand. ■

Donc avec le codage ε-typique le nombre de bits (par symbole) est borné supérieurement en gros par l'entropie de la probabilité des symboles. Peut-on faire mieux? cela serait le cas si on pouvait montrer que les observations se concentrent "encore plus" dans des sous-ensembles des ensembles typiques... Or, a priori, on a montré que justement la probabilité au sein des ensembles typiques est quasi-uniforme donc ça semble assez compromis. C'est ce que l'on va voir.

6.6 Les ensembles typiques sont "optimaux"

En quelque sorte, on va montrer que les ensembles typiques sont les bons objets, non seulement les observations vont s'y concentrer, mais en plus on ne peut pas espérer

mieux. Soit B_δ^n le *plus petit ensemble* tq.

$$\mathbb{P}(X \in B_\delta^n) \geq 1 - \delta \tag{181}$$

peut-il être de taille plus petite que l'ensemble typique? La réponse est non, et cela est dû au théorème suivant

Théorème 15 *(**optimalité des ens. typiques**)*
Avec $X = (x_1, \ldots, x_n)$ où les x_i sont des v.a iid de loi $p(x)$ $\forall \delta, \delta' > 0$,

$$\mathbb{P}(X \in B_\delta^n) \geq 1 - \delta \Rightarrow \frac{1}{n} \log_2 |B_\delta^n| \geq \mathbb{H}(x) - \delta' \tag{182}$$

Démonstration 15. La démonstration se focalise sur l'intersection entre B_δ^n et T_n^ε.

$$\mathbb{P}(T_n^\varepsilon \cap B_\delta^n) = \mathbb{P}(T_n^\varepsilon) + \mathbb{P}(B_\delta^n) - \mathbb{P}(T_n^\varepsilon \cup B_\delta^n) \geq (1 - \varepsilon) + (1 - \delta) - 1 = 1 - \varepsilon - \delta \tag{183}$$

Or, tout élément de l'intersection est donc élément de T_n^ε et l'on peut se servir des inégalités 171. Il vient

$$\mathbb{P}(T_n^\varepsilon \cap B_\delta^n) = \sum_{X \in T_n^\varepsilon \cap B_\delta^n} p(X) \leq |T_n^\varepsilon \cap B_\delta^n| \times 2^{-n(\mathbb{H}(x)-\varepsilon)} \leq |B_\delta^n| \times 2^{-n(\mathbb{H}(x)-\varepsilon)} \tag{184}$$

Qui se réécrit

$$|B_\delta^n| \times 2^{-n(\mathbb{H}(x)-\varepsilon)} \geq 1 - \varepsilon - \delta \Rightarrow \frac{1}{n} \log_2 |B_\delta^n| \geq \mathbb{H}(x) - \varepsilon + \frac{1}{n} \log_2(1 - \varepsilon - \delta) \tag{185}$$

On peut alors pour tout δ et δ' trouver ε et n suffisamment grand [66] pour que

$$\frac{1}{n} \log_2 |B_\delta^n| \geq \mathbb{H}(x) - \delta' \tag{186}$$

■

Donc, l'ensemble de taille minimum qui concentre le lieu des observations est bien l'en-

66. Par ex. $\varepsilon \leq \delta'/2$ et $n \geq \log_2(1 - \delta - \delta'/2)/\delta'/2$.

semble 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 14 est optimale***.

La géométrie certes nous donne la vision des ensembles typiques et le lien avec l'entropie, mais le code typique n'est pas du tout pratique. La raison est simple à comprendre, dans le code typique il est nécessaire de lever un bit si une séquence appartient à l'ensemble typique, mais encore faudrait-il être capable de tester si c'est le cas! Or, directement ce n'est pas praticable. Il nous faut donc trouver des moyens d'implémenter cette notion de codage typique dans ***des algorithmes efficaces qui atteignent la borne de Shannon***. On les appellent des ***codages entropiques instantanés***.

Pour aborder le sujet qui sera développé la fois prochaine, prenons une séquence $X = (x_1, x_2, \ldots, x_n)$ où chaque x_i prend sa valeur dans un alphabet $\mathcal{A} = \{a_k]\}_{1 \leq k \leq K}$. Soit donc a_k une valeur prise, celle-ci est codée par un mot binaire $w(k)$ de longueur $\ell(k)$, et définir un code c'est donner à chaque symbole a_k le mot $w(k)$ de manière à ce que la longueur soit minimum. On code non pas la séquence X tout entière, mais chacun de ses éléments. Ainsi, on veut minimiser la quantité

$$R = \sum_k \ell(k) p(a_k) \tag{187}$$

Une possibilité serait par exemple de ranger les symboles par ordre décroissant d'occurrence, et dans le cas de 4 symboles de définir le code à longueur variable suivant $w_1 = 0, w_2 = 10, w_3 = 101, w_4 = 111$, avec en arrière fond l'usage d'un mot binaire court (long) pour des symboles fréquents (rares). Deux remarques viennent immédiatement: *primo*, il faut connaitre *a priori* les probabilités $p(a_k)$, et *secundo* le code doit être décodable. Mettons que par une analyse préalable on ait une idée des fréquences d'occurrence des symboles qui nous donne une première approximation des vraies probabilités, le second problème est plus sérieux. En effet si j'envoie $(a_2 a_2)$ via le code 1010, le récepteur peut comprendre certes $(a_2 a_2)$ mais aussi $(a_3 a_1)$ ce qui est particulièrement désagréable, le code n'est pas satisfaisant car il n'est pas décodable (sans ambiguïté). Notons, que le récepteur qui reçoit 1111 par une simple erreur de transmission commencerait à décoder a_4 puis ne serait pas quoi faire du 1 restant, il attendrait sans doute des bits supplémentaires...

Mettons que le canal de communication soit idéal, l'origine de l'ambiguïté entre

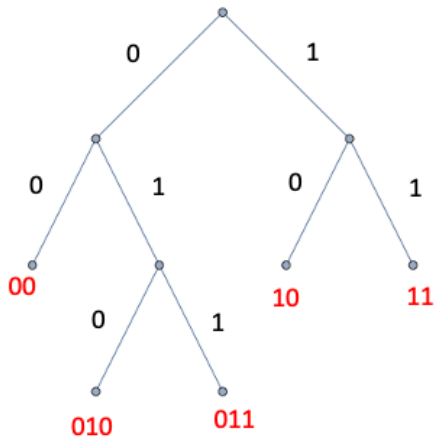

FIGURE 17 – Représentation d'un arbre binaire pour obtenir un codage satisfaisant la contrainte de préfixe en prenant les feuilles de l'arbre.

a_2a_2 et a_3a_1 réside dans le fait que le mot w_2 est le début du mot w_3. Mettre des séparateurs n'aiderait pas car il faudrait coder le séparateur ce qui rallongerait l'information à transmettre. Il nous faut plutôt une contrainte dite de ***préfixe*** qui stipule qu'***aucun mot binaire est le début d'un autre mot***. Avec ce type de condition, d'une certaine manière, nous avons un séparateur sans coût additionnel d'augmentation du nombre de bits à transmettre. Cependant, ce préfixage est une contrainte, donc il serait utile de pouvoir construire des codages simplement. Là la remarque élégante à faire c'est ***la correspondance avec un arbre binaire***, et ***la contrainte de préfixage est satisfaite ssi on prend les feuilles de l'arbre*** (Fig. 17). En effet, si on prend uniquement les feuilles de l'arbre, on ne peut avoir deux mots dont l'un est le commencement de l'autre, et inversement supposons que j'ai un code satisfaisant la condition de préfixage, alors je peux construire l'arbre et le couper au niveau des mots codes qui sont de facto les feuilles de l'arbre.

Le point qui nous reste à voir, sans parler des problèmes de canal bruité et/ou défectueux, c'est l'optimisation de R. Cependant, $\ell(k)$ correspond à la profondeur de $w(k)$ dans l'arbre binaire. Donc, R ***représente la profondeur moyenne des feuilles de l'arbre binaire représentant le code***. Ainsi, ***le problème est supposant connues les*** $p(a_k)$ ***comment construire l'arbre binaire dont les feuilles ont en moyenne la profondeur la plus petite possible***. Incidemment, cela valide aussi le fait de prendre des mots courts

pour les séquences les plus probables. La réponse donne le code optimum.

7. Séance du 23 Févr.

Pour mémoire, nous avons vu à la dernière séance que si l'on veut coder une série de valeurs $X = (x_1, x_2, \ldots, x_n)$ où les x_i sont des éléments d'un alphabet de taille fini $\mathcal{A} = \{a_k\}_{1 \leq k \leq K}$, alors le nombre moyen de bits par symbole satisfait [67]

$$\mathbb{H}(x) \leq R \leq \mathbb{H}(x) + C\varepsilon \tag{188}$$

pour un codage ε***-typique*** (Déf. 11), et que l'on ne peut pas faire mieux. D'un point de vue géométrique, le codage des n symboles qui est un élément de $\mathcal{A}^n$ est en fait un élément de l'ensemble typique T_n^ε qui est de taille environ $2^{n\mathbb{H}(x)}$, et ***donc le nombre de bits est de l'ordre du log de cette quantité soit*** $\mathbb{H}$ ***par symbole***. Et finalement, il n'existe pas d'ensemble plus petit que T_n^ε.

7.1 Codage instantané (1 symbole à la fois)

La mise en pratique d'un codage typique n'est pas possible en général, car il faut pouvoir connaitre si une séquence est élément de l'ensemble typique ou non pour lever un bit. On a vu que l'on peut envisager des codes plus simples (dit instantanés) qui opèrent symbole par symbole, en assignant 1 mot binaire w_k à chaque symbole a_k satisfaisant à une ***contrainte de préfixage***, et que cela peut se réaliser en construisant ***un arbre binaire*** où les mots codes en sont les ***feuilles*** (Fig. 17). Chaque mot w_k a une longueur ℓ_k, et donc le problème devient: connaissant les probabilités d'occurrence $p(a_k) = p_k$ de chaque symbole, trouver les w_k tels que *primo* le décodage soit possible (satisfait par le préfixage)

67. NDJE: la borne supérieure est celle du théorème 14 obtenu pour les ensembles typiques, tandis que la borne inférieure n'a pas été démontrée en fait. On peut dégager les grandes lignes d'une démonstration: la relation entre taille de l'ensemble typique et l'entropie est en gros $\mathbb{H}(x) \approx 1/n \log_2 |T_n^\varepsilon|$ qui en d'autres termes est le nombre de bits moyen pour coder les éléments de l'ensemble. Si on prend en compte le résultat Sec. 6.6 (postérieur au Th. 14) qui stipule qu'il n'y a pas d'ensemble qui concentre mieux l'information que l'ensemble typique, alors la valeur de R ne peut être plus petite que l'entropie $\mathbb{H}(x)$. Ceci dit la démonstration précise dans le cadre des ensemble typique reste à établir. Je présente une démonstration établie avec le lemme de Kraft dans cette séance.

et *secundo* la longueur moyenne d'un symbole codé (nombre de bits moyen)

$$R = \sum_k \ell_k p_k \tag{189}$$

soit la plus petite possible. Notons que la longueur ℓ_k correspond exactement à la profondeur du mot w_k dans l'arbre binaire.

Intuitivement, pour construire l'arbre, on se dit que les mots les plus fréquents doivent être codés avec des mots courts donc avec des feuilles de l'arbre proche de la racine. Et inversement pour les mots les moins fréquents sont codés avec des mots correspondant à des feuilles loin de la racine. Si ceci est en arrière plan pour construire l'arbre, il nous faut comprendre le lien entre ℓ_k et p_k. Pour cela nous allons voir un premier théorème de C. Shannon.

Théorème 16 *(**Code de Shannon**)*
Soit une source X de symboles a_k dont les probabilités sont connues et notées p_k, alors pour un code de préfixe nous avons

$$R \geq \mathbb{H} = -\sum_k p_k \log_2 p_k \tag{190}$$

et il existe un code dit de Shannon tq.

$$R \leq \mathbb{H} + 1 \tag{191}$$

C'est-à-dire que l'inefficacité est d'au plus 1-bit par rapport à l'entropie.

La démonstration est basée sur un lemme très important en Théorie de l'Information qui est le suivant [68]

Lemme 1 *(**Kraft**)*
Tout code de préfixe avec K mots binaires w_k de longueur $\ell(w_k) = \ell_k$ satisfait

68. de Leon Gordon Kraft, lemme qu'il publie dans sa thèse en 1949.

l'inégalité suivante:

$$\sum_{k=1}^{K} 2^{-\ell_k} \leq 1 \qquad (I1) \qquad (192)$$

Réciproquement, si la collection de ℓ_k satisfait l'inégalité $(I1)$ alors il existe un code de préfixe $\{w_k\}_{1\leq k\leq K}$ tel que les longueurs des mots binaires satisfont $\ell(w_k) = \ell_k$.

Démonstration 1. Prenons la condition nécessaire. Supposons que j'ai un code de préfixe (arbre binaire dont les feuilles sont les mots codes), et soit la profondeur maximale de l'arbre:

$$m := \max_k \ell_k \qquad (193)$$

Pour chaque feuille de l'arbre binaire original, on la fait devenir la racine d'un nouvel arbre binaire que l'on déploie pour atteindre la profondeur m. Pour une feuille originale de profondeur ℓ_k, le nombre de feuilles de son arbre à la profondeur m est de $2^{m-\ell_k}$ (Fig. 18). A cette profondeur maximale, tous les mots sont disjoints, et leur nombre est inférieur au nombre total de mots possibles à cette profondeur (à cause de la présence des feuilles originales à cette profondeur)

$$\sum_{k=1}^{K} 2^{m-\ell_k} \leq 2^m \qquad (194)$$

Ce qui donne la relation $(I1)$ en divisant de part et d'autre part 2^m.

Inversement, on a une famille de mots de longueur ℓ_k. On commence par les ordonner: $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_K$ (moyennant une redéfinition des indices). Par exemple dans la figure 18 les profondeurs des feuilles (vertes) de gauche à droite se lisent $\{3, 3, 5, 5, 5, 4, 4, 1\}$ que l'on peut réarranger selon $\{1, 3, 3, 4, 4, 5, 5, 5\}$. Ensuite, on lit la profondeur maximale donnant $m = 5$, on construit alors l'arbre complet jusqu'à cette profondeur, puis on attache les K sous-arbres de tailles $2^{m-\ell_k}$ en partant par exemple de la gauche. On sait que l'on peut les inclure grâce à l'inégalité $(I1)$ (Fig. 19). Les racines des sous-arbres sont identifiées comme les feuilles de l'arbre binaire de codage, et l'on constate qu'elles forment alors un code à préfixe dont les longueurs $\ell(w_k)$ sont bien les ℓ_k. ■

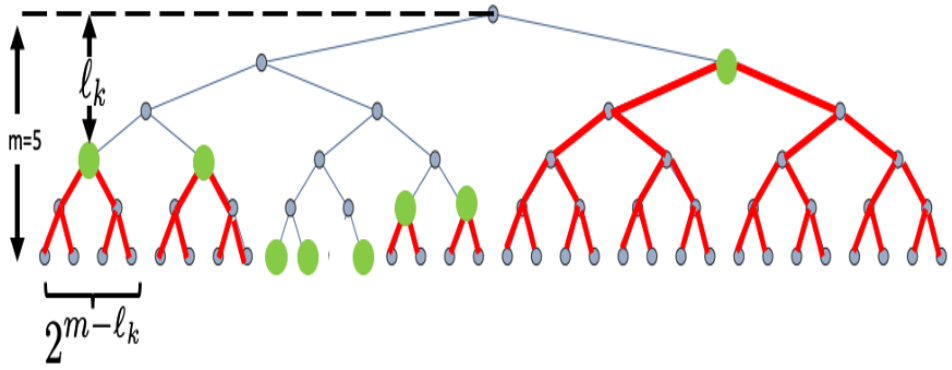

FIGURE 18 – Exemple d'un arbre binaire de profondeur maximale m, et dont les feuilles (points verts) sont prolongés par des sous-arbres binaires jusqu'à la profondeur maximale. Le lemme de Kraft donne une contrainte sur le nombre total de feuilles existantes à la profondeur maximale.

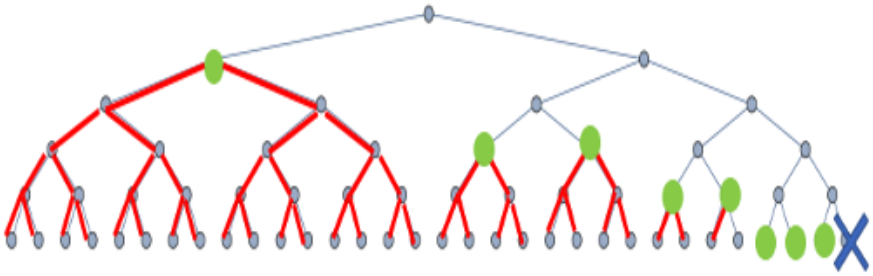

FIGURE 19 – Reconstruction d'un arbre binaire à partir des longueurs des mots (lemme de Kraft). La croix signifie in fine que l'on supprime une feuille non utilisée.

Revenons à la démonstration du théorème 16.

Démonstration 16. Au problème posé de trouver les mots w_k tels que leurs longueurs ℓ_k minimisent

$$R = \sum_{k=1}^{K} \ell_k p_k \tag{195}$$

on y ajoute la condition

$$\sum_{k=1}^{K} 2^{-\ell_k} \leq 1 \tag{196}$$

pour satisfaire la condition de préfixage. On a alors un problème de minimisation linéaire avec contrainte convexe. La solution est unique et donnée par les multiplicateurs de Lagrange [69]. On se donne alors le lagrangien

$$\mathscr{L}(\{\ell_k\}, \lambda) = \sum_{k=1}^{K} \ell_k p_k + \lambda \left(\sum_{k=1}^{K} 2^{-\ell_k} - 1 \right) \tag{197}$$

Le point selle (ou col) satisfait $\forall i$

$$\frac{\partial \mathscr{L}}{\partial \ell_i} = p_i - \lambda \; 2^{-\ell_i} \; \log_e 2 = 0 \tag{198}$$

En sommant sur l'ensemble des i, la somme des probabilités vaut 1 et la contrainte se transformant en égalité fournit alors

$$\lambda^* \log_e 2 = 1 \qquad \boxed{\ell_i^* = -\log_2 p_i} \tag{199}$$

On a établit de ce fait le lien entre la longueur du mot code ℓ_k et la probabilité p_k d'occurrence du symbole a_k.

Le set de valeurs minimales $\{\ell_k^*\}_k$ fournit la valeur correspondante de R:

$$R_{min} = -\sum_k p_k \log_2 p_k = \mathbb{H} \tag{200}$$

NDJE: Avec le lemme de Kraft, on peut démontrer que pour un code quelconque à préfixe $R \geq \mathbb{H}$. En effet, soit un code à préfixe basé sur les probabilités $(p_k)_k$ et le jeu de

69. Voir Cours 2018 Sec. 8.3.

longueur ℓ'_k. Le lemme de Kraft exige alors que

$$C' = \sum_{k=1}^{K} 2^{-\ell'_k} \leq 1$$

Alors, définissons les probabilités $(q_k)_k$ selon

$$q_k := \frac{2^{-\ell'_k}}{C'} \Rightarrow -\log_2 q_k = \ell'_k + \log_2 C'$$

Or,

$$D(p\|q) = \sum_k p_k \log p_k - \sum_k p_k \log q_k \geq 0$$

qui donne

$$R' \geq H(x) - \log C' \geq H(x)$$

Donc, pour n'importe quel code binaire à préfixe R est supérieur à $\mathbb{H}$, et **le résultat du théorème nous dit alors que l'on pourrait atteindre possiblement la borne en prenant les longueurs $\{\ell_k^*\}_k$.**

Mais pourquoi en pratique ce n'est pas le cas, d'où la présence de la seconde inégalité du Th. de Shannon? La raison est simple, les ℓ_i^* doivent représenter des profondeurs dans un arbre, donc ce sont des nombres entiers, or **les p_k ne sont pas forcément des puissances de 2**. On a donc en pratique une approximation de ce codage optimal (dit codage entropique) dont Shannon donne l'exemple suivant

$$\boxed{\tilde{\ell}_k = \lceil -\log_2 p_k \rceil} \qquad (Shannon) \tag{201}$$

qui satisfait l'inégalité de Kraft car $\lceil x \rceil \geq x$. Concernant la valeur de R, on a comme $\lceil x \rceil \leq x + 1$

$$R \leq \sum_k p_k(-\log_2 p_k + 1) = \mathbb{H} + 1 \tag{202}$$

ce qui donne la seconde inégalité de Shannon. ■

Ce théorème est remarquable car il est constructif et l'on a les bornes inf. et sup. sur le

nombre de bits moyen par symbole. Cependant, il n'est pas optimal, il nous fait juste le lien entre un code théorique et un code réalisable. *Est-ce grave docteur que ce code ne soit pas optimal?* Ben, tout dépend du problème, mais si l'on considère des images, la valeur du pixel est certes codée par example sur 8 bits (0: noir, 255: blanc), mais en réalité en moyenne, on a plutôt 1/4 de bit par pixel, donc dans ce cas ***rajouter 1 bit par pixel dû à l'inefficacité du code est très pénalisant***. Donc, il faut se donner la peine de réduire la borne pour que typiquement

$$R \leq \mathbb{H} + O(\varepsilon) \tag{203}$$

L'inefficacité vient de ce que les longueurs des mots code sont contraintes à être entières. On va alors s'inspirer du codage typique à savoir considérer non plus 1 seul symbole à la fois, mais ***des blocs de*** n ***symboles*** surtout que si n est grand, on sait que l'on va concentrer les probabilités. On s'attend à perdre 1 bit par bloc de taille n, donc par symbole on ne perdra plus que $1/n$ bit.

7.2 Codage entropique par bloc

Soit donc $X = (x_1, \ldots, x_n) \in \mathcal{A}^n$, chaque x_i est toujours élément d'un alphabet de K symboles, donc l'ensemble des X est de taille $|\mathcal{A}^n| = K^n$. Si on applique le théorème de la section précédente, on obtient alors que le nombre de bits moyen pour coder X selon le code de Shannon satisfait

$$\mathbb{H}(X) \leq R_X \leq \mathbb{H}(X) + 1 \tag{204}$$

Si on fait l'hypothèse que les $(x_i)_i$ sont *iid* alors naturellement

$$\mathbb{H}(X) = n\mathbb{H}(x) \tag{205}$$

Au passage, quand on considère les variables *iid* on se place dans le cas le plus défavorable. En effet, si tel n'est pas le cas l'entropie réelle est plus petite que l'entropie *iid*[70], et donc le code est moins efficace que si on avait pris en compte la corrélation entre les symboles. Nous y reviendrons. Mais restons dans le cas *iid*, alors le nombre de bit par symbole

70. NDJE: pour le dire avec un arrière fond de méca. stat. : "il y a moins de désordre à cause des corrélations".

satisfait

$$\mathbb{H}(x) \leq R \leq \mathbb{H}(x) + \frac{1}{n} \tag{206}$$

Donc, on a en main un algorithme qui en principe devient optimal quand on considère des blocs de symboles de plus en plus grand (n tendant vers l'infini). Cependant, la solution de Shannon n'est pas optimale car pour n fixé il ne garantit pas que l'on ait en main le codage qui réalise le minimum de R.

7.3 Code optimal de Huffman

L'idée derrière le code optimal vient de la réflexion suivante: si l'on considère l'arbre associé au code optimal à préfixe, alors une feuille plus profonde à toujours une probabilité plus faible que celle d'une feuille moins profonde. Autrement dit, plus on s'enfonce dans l'arbre plus on rencontre des symboles moins probables. En effet, considérons la situation de la figure 20. Si toute chose égale par ailleurs $p_{k'} \leq p_k$ pour deux feuilles dont $\ell_{k'} \leq \ell_k$ alors [71]

$$p_{k'}\ell_k + p_k\ell_{k'} \leq p_k\ell_k + p_{k'}\ell_{k'} \tag{207}$$

Donc en échangeant les symboles k et k' on obtient un codage dont le R est meilleur. David Albert Huffman (1925-99) au MIT en 1951 a utilisé cette idée et résolu le problème posé par son professeur Rober Fano [72] (1917-2016) qui lui-même avait mis au point l'arbre en partant des probabilités les plus grandes (top-bottom) tandis que Huffman procède en partant des probabilités les plus petites (bottom-up).

Définition 12 ***(Code de Huffman)***
Supposons que l'on range les K symboles selon les probabilités croissantes: $p(a_k) \leq p(a_{k+1})$. On va relier le problème à K symboles à celui sur $K-1$ symboles. Pour ce faire, on part des 2 symboles les moins fréquents (a_1, a_2) pour constituer un symbole $a_{1,2}$ (a_1 ou a_2) dont la probabilité est la somme $p_1+p_2 = p_{1,2}$. En éliminant (a_1, a_2) au profit de $a_{1,2}$ on réduit le problème de 1 unité passant de K à $K-1$ symboles. Ainsi, si l'on dispose d'un code de taille optimale pour les $K-1$ symboles $\{a_k\}_{k>2} \cup \{a_{1,2}\}$,

71. NDJE: il suffit de se rendre compte que $(p_k - p_{k'})(\ell_k - \ell_{k'}) \geq 0$ et de développer cette expression.
72. Son frère ainé Ugo Fano est bien connu des physiciens nucléaires.

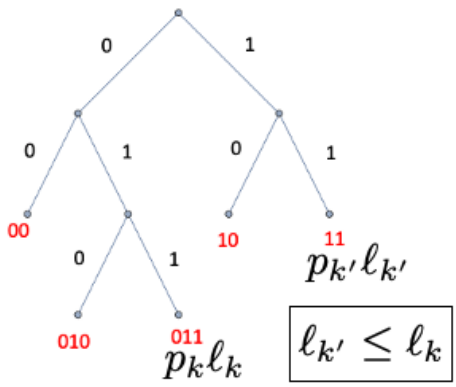

FIGURE 20 – Situation où une feuille plus profonde à une probabilité plus grande, $p_k \geq p_{k'}$: échanger les symboles concernés permet de diminuer la valeur de R.

> *alors on dispose d'un code optimal pour les K symboles en divisant la feuille $a_{1,2}$ en deux sous-feuilles.*

La démonstration est basée sur la réflexion faite ci-dessus sur la position relative des probabilités dans l'arbre, et en faisant également la réflexion que pour un arbre complet à préfixe il n'y a jamais une feuille unique dont le mot code a la longueur maximale. Un exemple d'implémentation du code de Huffman est donné dans un notebook Python associé à ce cours [73]. ***Le code de Huffman est réellement optimal au sens pratique, c'est-à-dire que l'on ne peut pas faire mieux (à moins de faire un codage par blocs), et son inefficacité par rapport au code entropique est largement inférieur à 1 bit***. Dans l'exercice proposé dans le notebook vous pourrez constater que $R \approx 2.24$ tandis que l'entropie est de 2.18 soit une inefficacité de 0.06 bits. Cependant, le code de Huffman a quelques défauts dont la nécessité de transmettre l'arbre de codage, et la rigidité du codage qui doit être recalculé pour chaque texte transmis.

7.4 Entropie différentielle

Tous les développements algorithmiques précédents nous résolvent parfaitement le cas d'alphabet de taille fini. Le problème en pratique quand on dispose de mesures, c'est

73. https://github.com/jecampagne/cours_mallat_cdf/cours2022, `Simple_huffman_code.ipynb`.

que celles-ci sont des ***nombres réels***. Donc, il va falloir se définir les équivalents d'information, entropie et les algorithmes associés sachant que possiblement un réel est représenté par un nombre infini de bits. Cependant, dans le monde réel si l'on peut dire, les *"floats"* sont représentés par 32, 64-bits ou plus parfois, et donc on passe dans le domaine fini au prix d'une quantification qui introduit une erreur. Avant d'envisager cela, on va voir comment on étend la théorie, entropie et ensemble typique, aux valeurs réelles et par ce biais nous allons poser un pont avec la première partie du cours à savoir l'information de Fisher.

Définition 13 *(**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

$$\boxed{\mathbb{H}_d = -\int p(x) \log p(x)\ dx} \tag{208}$$

Contrairement à son équivalent "discret" (Déf. 6), l'entropie différentielle n'est pas forcément positive. Voici un exemple:

$$X \sim \mathcal{U}([0, a]) \Rightarrow \mathbb{H}_d = \log a \tag{209}$$

donc si $a < 1$, entropie différentielle est négative. Il faut voir cette entropie comme relative à une mesure de référence ici celle de Lebesgue (dx). Concernant la distribution gaussienne, on a

$$X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \mathbb{H}_d = \frac{1}{2} + \frac{1}{2}\log\left(2\pi\sigma^2\right) = \frac{1}{2}\log(2\pi e) + \log\sigma \tag{210}$$

En fait, a ou σ peuvent être vus comme des facteurs d'échelle, et d'une manière générale ($\forall \alpha > 0$)

$$\boxed{\mathbb{H}_d(\alpha X) = \mathbb{H}_d(X) + \log\alpha} \tag{211}$$

qui vient du fait que $p(\alpha x)(\alpha dx) = p(x)dx$, soit $p(\alpha x) = \frac{1}{\alpha}p(x)$. Donc, l'intervalle de mesure donne le caractère relatif à la définition de l'entropie.

A partir de cette définition d'entropie, on peut étendre toutes les notions vues, et en particulier les ensembles typiques. Ce que l'on va vérifier c'est que si on prend n *v.a*

iid, la probabilité jointe satisfait

$$p(x_1, \ldots, x_n) = \prod_i p(x_i) \approx 2^{-n\mathbb{H}_d(x)} \tag{212}$$

Ceci est le pendant de l'équation 171 de la section 6.4 sur les ensembles typiques sur un alphabet de taille finie. En fait, nous avons la propriété suivante

$$-\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) \xrightarrow[n\to\infty]{prob.} -\mathbb{E}_{x\sim p(x)}[\log p(x)] = \mathbb{H}_d(x) \tag{213}$$

Ainsi, la log-probabilité d'un bloc de n *v.a iid* se concentre à un epsilon près autour de sa moyenne, c'est-à-dire l'entropie différentielle. On a alors envie de définir les ensembles qui contiennent presque toutes les réalisations de ces n-blocs, ce sont les ensembles typiques associés. La définition suit celle donnée par l'expression 169 en utilisant l'entropie différentielle [74]:

$$\boxed{T_n^\varepsilon = \left\{\{x\} \in \mathbb{R}^n, \left|-\frac{1}{n}\log p(\{x\}) - \mathbb{H}_d(x)\right| \leq \varepsilon\right\}} \tag{214}$$

Et la convergence en probabilité, pour n assez grand nous garantie que

$$\boxed{\mathbb{P}(\{x\} \in T_n^\varepsilon) \geq 1 - \varepsilon} \tag{215}$$

Si le parallèle est frappant, quelle est la différence entre les deux versions (*discrète* vs *continue*) de la théorie? En *discret*, l'entropie (toujours positive) mesure à ε prés le nombre de bits qu'il faut pour coder un élément de l'ensemble typique (Voir le code typique Déf. 11), ou dit autrement cela donne le nombre d'éléments dans l'ensemble typique (Prop. 6). Dans le cas *continue*, non seulement l'entropie différentielle n'est pas garantie d'être positive, mais ***le nombre d'éléments d'un ensemble typique est infini***. Le lien est donné par le théorème suivant

74. NDJE: dans l'expression 169 et les suivantes, $\{x\} \in \mathcal{A}^n$ contrairement à ce qui a pu être écrit dans des versions antérieures. Je vous prie de m'en excuser.

Théorème 17 *(**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

$$\boxed{(1-\varepsilon)2^{n(\mathbb{H}_d(x)-\varepsilon)} \leq V(T_n^{\varepsilon}) \leq 2^{n(\mathbb{H}_d(x)+\varepsilon)}} \tag{216}$$

Démonstration 17. On ne donne qu'une partie de la démonstration qui calque son homologue en discret. Notant $\{x\} = X$, l'appartenance de X à T_n^{ε} signifie que

$$2^{-n(\mathbb{H}_d(x)+\varepsilon)} \leq p(X) \leq 2^{-n(\mathbb{H}_d(x)-\varepsilon)} \tag{217}$$

Or,

$$1 = \int p(X)dX \geq \int_{T_n^{\varepsilon}} p(X)dX \geq 2^{-n(\mathbb{H}_d(x)+\varepsilon)} \int_{T_n^{\varepsilon}} dX \tag{218}$$

ce qui donne un coté de la double inégalité. ■

Ainsi, les volumes typiques ont des volumes tels que

$$V(T_n^{\varepsilon}) \approx 2^{n\mathbb{H}_d(x)} \approx \frac{1}{p(X)} \tag{219}$$

et donc la probabilité est quasi-constante donnée par l'inverse du volume. L'entropie différentielle peut être vue alors comme le log en base 2 de la longueur du coté d'un pavé de volume équivalent (en dimension n).

Pour utiliser des codes discrets comme celui de Huffman, il va nous falloir mettre en place une sorte de pavage des ensembles typiques T_n^{ε}. Chaque boule de ce pavage définit un symbole et l'ensemble des symboles permet de décrire n'importe quel élément de T_n^{ε} à une (petite) erreur prés. La difficulté est de trouver des pavages "optimaux". Nous verrons qu'il y a des façons plus simples d'appréhender le problème comme dans le cas discret. Avant cela, nous allons faire le lien entre cette notion d'entropie et l'inférence (de Fisher).

7.5 Principe d'entropie maximum

NDJE: Afin de refonder la Mécanique Statistique, surtout dans le but d'attaquer des problèmes hors d'équilibre, un principe[75] *a été élaboré en 1957 par Edwin Thompson Jaynes (1922-98): il s'agit du* ***Principe d'Entropie Maximale****. Il se trouve qu'à partir de ce principe, on peut retrouver toute la Mécanique Statistique en la considérant comme une théorie déductive (ie. théorie de l'inférence), car on retrouve naturellement la fonction de partition de Gibbs. Ainsi, l'entropie de Shannon doit être prise selon E. Jaynes comme le concept premier dont on découle les autres observables.*

L'idée de Jaynes est qu'il s'agit de savoir comment utiliser au mieux ***les informations partielles*** ou ***contraintes*** que l'on a sur un système. Mettons par exemple que l'on s'intéresse à un gaz dont la température est fixée, alors selon la Théorie de Boltzmann, le système va "optimiser" sa configuration de telle façon que la probabilité de la configuration est donnée par

$$P \approx Z^{-1} \exp\Big\{-\frac{\mathcal{H}}{k_B T}\Big\} \tag{220}$$

avec $\mathcal{H}$ l'hamiltonien du système (égal à l'énergie totale constante) qui régit les équations du mouvement de chaque particules de gaz. On voit là un lien avec l'idée de l'ensemble typique, où la configuration du système est 1 point de l'ensemble et la probabilité y est quasi-constante. ***Les ensembles typiques sont les plus grands ensembles correspondant à la contrainte de température fixée***. Cette idée Jaynes va l'étendre au-delà de la Mécanique Statistique.

Jaynes fait remarquer que dans beaucoup de problèmes, ***les observables dont on dispose sont des valeurs moyennes***. Ainsi, on peut écrire pour une observable $U_k(x)$ $(k \leq K)$ que l'on a accès à

$$\int p(x) U_k(x) dx = \mathbb{E}_{x\sim p}[U_k(x)] = \mu_k \tag{221}$$

mais on ne connait que μ_k alors que $p(x)$ est ici l'inconnue. Les fonctions $U_k(x)$ peuvent être plus ou moins compliquées. Disposant des $(\mu_k)_k$ la question est alors: ***quelle densité de probabilité*** $p(x)$ ***sous-jacente aux processus étudié va d'une manière "naturelle"***

75. au sens original, il s'agit d'un guide de mise en ordre du monde, certains y verront un axiome/postulat.

satisfaire ces contraintes d'observations? Or, a priori se donner uniquement les $(\mu_k)_k$ n'est pas suffisant, il nous faut un principe directeur. L'idée est de contraindre $p(x)$ à être ***la plus uniforme possible dans un espace de volume maximum***. Voilà le lien avec ***les ensembles typiques***. Or, maximiser le volume, signifie ***maximiser l'entropie différentielle***. Jaynes cherche ***une distribution de probabilité a priori la moins informative possible***[76]. Finalement, on a un problème d'optimisation avec contraintes convexes qui donne le théorème de Boltzmann/Gibbs de Physique Statistique suivant:

Théorème 18 ***(Boltzmann/Gibbs)***
Soit le problème de trouver la probabilité $p^*(x)$ telle que la fonction

$$H(p) = -\int p(x)\log p(x)dx$$

ainsi que K fonctions c_k de $\mathbb{R}^n \to \mathbb{R}$ satisfont

$$p^* = \underset{p}{\operatorname{argmax}}\ H(p); \qquad \text{et} \quad \forall k, c_k(p) = 0 \tag{222}$$

Si la solution p^* existe, elle est unique et s'écrit

$$\boxed{p^*(x;\theta) = Z^{-1}\exp\Big\{-\sum_k \theta_k U_k(x)\Big\}} \tag{223}$$

avec $\theta = (\theta_k)_k$ les multiplicateurs de Lagrange. On retrouve la densité de probabilité paramétrée de Fisher.

De plus, on sait que $\mathbb{H}(p^*) \geq \mathbb{H}(p_{vraie})$ mais si elles sont égales alors $p^* = p_{vraie}$. Ce résultat est relié au problème inverse de trouver les $(U_k)_k$ pour approximer la vraie probabilité, ce qui est en lien avec les problématiques d'architecture de réseaux de neurones.

Démonstration 18. Envisageons la première partie du théorème. La solution réalise l'ex-

76. NDJE: On peut noter que E. Jaynes s'inscrit dans la tradition subjectiviste des probabilités en suivant Harold Jeffreys (1891-1989) dans l'étude des *priors* non-informatifs.

tremum du lagrangien

$$\mathscr{L}(p,\theta) = H(p) + \sum_{k=1}^{K} \theta_k c_k(p) + \theta_0(\int p(x)dx - 1) \tag{224}$$

avec $c_k(x) = \mu_k - \int p(x)U_k(x)dx$. où les variables $(\theta_k)_{k\leq K}$ sont les multiplicateurs de Lagrange. Ainsi, p^* satisfait (dérivée au sens de Gâteaux)

$$\frac{\partial \mathscr{L}}{\partial p(x)} = -\log(p(x)) - 1 - \sum_{k=1}^{K} \theta_k U_k(x) + \theta_0 = 0 \tag{225}$$

Donc,

$$p^*(x) = \exp\left\{\theta_0 - 1 - \sum_{k=1}^{K} \theta_k U_k(x)\right\} \tag{226}$$

et la condition de normalité donne la valeur de θ_0 que l'on traduit par la fonction de partitions Z telle que

$$p^*(x;\theta) = Z^{-1} \exp\left\{-\sum_{k=1}^{K} \theta_k U_k(x)\right\} \qquad Z(\theta) = \int \exp\left\{-\sum_{k=1}^{K} \theta_k U_k(x)\right\} dx \tag{227}$$

Les θ_k sont fixés par les contraintes sur les moyennes

$$\int p^*(x;\theta)U_k(x)dx = \mu_k \tag{228}$$

■

Donc, si p^* existe, on a son expression. Cependant, ce n'est pas toujours le cas. Mais d'abord prenons un exemple classique où l'on a comme contrainte, la moyenne et la variance ou la matrice de covariance en dimension quelconque $\mathbb{R}^n$. Ainsi, soient les contraintes

$$\mathbb{E}[X] = \mu \qquad\qquad \mathbb{E}[(X-\mu)(X-\mu)^T] = \Sigma \tag{229}$$

Dans le cas $d = 1$, $X \in \mathbb{R}$ 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 $U_1(x) = x$ et $U_2(x) = x^2$. 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 $p^*(x)$

prend la forme

$$p^*(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} = \mathcal{N}(\mu, \sigma^2) \tag{230}$$

Ce qui se généralise en dimension n par

$$p^*(x) = \frac{1}{\sqrt{(2\pi)^n \ \det\Sigma}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} \tag{231}$$

d'entropie donnée par

$$H_d = \frac{1}{2}\left(n + \log((2\pi)^n \ \det\Sigma)\right) \tag{232}$$

Notons que si on avait imposé (en dimension 1) les contraintes sur $\mathbb{E}(X^k)$ avec $k = 1, 2, 3$ alors on n'aurait pu trouver de solution à cause de la divergence de l'intégrale due à la présence du terme en $e^{-\theta_3 x^3}$. Donc, il nous faut prendre $\theta_3 = 0$, ce qui restreint le nombre de contraintes. La solution trouvée (la gaussienne) sera d'entropie plus grande que celle du problème initial à 3 contraintes. On a donc une borne supérieure de l'entropie donnée par l'entropie de la gaussienne, mais il n'y a aucune distribution physique qui peut atteindre l'entropie correspondante au problème des 3 contraintes. Cependant, on peut faire un développement perturbatif pour itérativement s'approcher de la solution.

7.6 Lien avec l'inférence

A partir du théorème en se donnant des contraintes et le Principe d'Entropie Maximale, on arrive à une distribution de probabilité paramétrée de type exponentielle. On peut voir le problème autrement en prenant le Maximum de Vraisemblance. Ainsi

$$\ell(\theta) = \log p_\theta(x) \tag{233}$$

qui dans le cas d'une famille d'exponentielles (Th. 5) donne $\forall k$

$$-\nabla_{\theta_k}\ell(\theta) = U_k(x) - \mathbb{E}_{x\sim p_\theta}[U_k(x)] \tag{234}$$

Et si on calcule pour $\theta = \theta^*$ réalisant le MLE alors (Eq. 115)

$$\mathbb{E}_{x\sim p_{\theta^*_k}}[\nabla_\theta \ell(\theta)] = 0 \tag{235}$$

Donc on en déduit que

$$\mathbb{E}_{x\sim p_{\theta^*}}[U_k(x)] = \mathbb{E}_{x\sim p_\theta}[U_k(x)] = \mu_k \tag{236}$$

ce qui est vrai en particulier pour le jeu de θ qui donne la vraie probabilité [77], laquelle fixe les valeurs des μ_k d'où sa présence dans l'expression ci-dessus. Donc, le MLE satisfait les contraintes sur les moyennes et a une forme exponentielle. D'où

> **Théorème 19** *La solution d'entropie maximum est l'estimateur de maximum de vraisemblance (MLE).*

MLE et Principe d'Entropie Maximale sont deux notions équivalentes. ***En d'autres termes, se donner comme but de déterminer une distribution qui maximise l'entropie la plus uniforme possible (la moins informative) qui est le Principe d'Entropie Maximale, revient à se donner un modèle paramétré exponentiel dont on maximise la vraisemblance.***

En particulier, pour déterminer les paramètres de Lagrange, on peut procéder par descente de gradients que nous avons vu Sec. 3.6.2. Mais de nouveau, on tombe sur les problèmes d'instabilité et de conditionnement du Hessien qui pour mémoire n'est rien d'autre que l'Information de Fisher $I(\theta^*)$. Notons au passage que l'étape de calcul GD s'écrit:

$$\theta_k^{t+1} = \theta_k^t + \varepsilon(\mathbb{E}_{\theta_t}(U_k) - \mu_k) \tag{237}$$

Or, le terme $\mathbb{E}_{\theta_t}(U_k)$ est difficile à calculer car il faut estimer cette intégrale

$$\mathbb{E}_{\theta_t}(U_k) = \int U_k(x) p_{\theta_t}(x) dx \tag{238}$$

Cela se fait par des méthodes d'échantillonnage Monte Carlo (Importance Sampling, Metropolis-Hastings, Gibbs Sampling, Markov Chain,...) qui demandent beaucoup de ressources et cela à chaque itération de GD.

Revenons à la 2nd propriété du théorème 18. L'observation est que si l'entropie de la solution trouvée est bien égale à celle de la distribution sous-jacente alors c'est la bonne solution (se remémorer le cas où l'on n'atteint pas l'entropie du problème). Ceci

77. NDJE: rappel dans le cadre de Fisher, la vraie probabilité est de la même famille que les p_θ.

est intéressant, si on retourne le problème. En Mécanique Statistique, on se donne les observables et on demande de calculer l'état du système, mais en ML ce n'est pas vraiment sous cette forme que le problème se pose. En effet, en ML en général on dispose d'observables distribuées selon une distribution $\bar{p}$ (inconnue), à partir desquelles on se donne des descripteurs (les $U_k(x)$) dont on calcule les valeurs moyennes (variances et autres moments). Ces descripteurs nous donnent un modèle de $p(x)$ que l'on espère s'approcher autant que possible de $\bar{p}(x)$. En fait, on cherche les "bons" $U_k(x)$ qui peuvent par exemple être le résultat de la chaine de cascades des couches d'un réseau de neurones. Ce que l'on connait, c'est l'entropie du système étudié $\mathbb{H}(\bar{p})$, et l'on sait que l'entropie du modèle est toujours plus grande ($\mathbb{H}(p) \geq \mathbb{H}(\bar{p})$). Donc, ***ce que l'on cherche alors c'est à minimiser l'entropie maximum*** (minimax) [78]. D'une manière générale, on se dit que l'on va définir ***une approximation en fixant la moyenne et la covariance*** ce qui correspondrait en Physique au terme d'énergie cinétique. Cependant, on sait maintenant que l'on aboutit à un modèle gaussien lequel n'est pas satisfaisant pour bon nombre de problèmes (Sec. 4.6). Il nous faut d'autres contraintes, mais comment les obtenir? Une méthode est de créer ***des représentations parcimonieuses***. Or dans ce cas, la distribution des coefficients dans la base (ex. ondelettes) d'une image (ex. Fig. 58 Cours 2021) n'est pas du tout gaussienne mais plutôt laplacienne, car la plupart des coefficients sont nuls et seulement quelques coefficients sont importants. Ainsi, le modèle gaussien ne va pas convenir en particulier on n'arrivera pas à comprimer l'image, ou le champ en 1D, 2D, 3D, etc. ***On impose alors que les moments des descripteurs reflètent cette parcimonie des coefficients.*** Ce sont les nouvelles contraintes recherchées.

8. Séance du 2 Mars

8.1 Vers la compression par transformée orthogonale

Nous allons aborder la théorie de Shannon par un coté plus applicatif en considérant le problème de ***la compression de signaux***, c'est-à-dire la réduction du nombre de bits pour représenter par exemple de l'image, de la vidéo, du son, etc. On a bien là deux aspects

78. NDJE: En Mécanique Statistique, c'est le problème d'énergie libre du modèle par rapport à l'énergie libre du système (ex. théorie du champ moyen).

à prendre en compte: le point de vue de la ***Théorie de l'Information*** et le point de vue de la ***Représentation***, thématique abordée en 2021. Pour faire court, que cela soit dans le cadre de Fisher, ou celui de Shannon, il y a une hypothèse fondamentale d'***indépendance des observations***. Mais, en pratique considérant par exemple des images ou de l'audio, il y a énormément de ***redondance*** et en même temps de la ***structuration*** fondamentale puisque c'est elle qui nous permet de reconnaitre un visage, une voix, etc. Et finalement, ***nous ne sommes pas face à des mesures indépendantes***. Donc, le problème qui se pose est de savoir ***comment utiliser la redondance, la structuration des observations afin de minimiser le nombre de bits pour les représenter/transmettre, et de se rapprocher d'une certaine façon de problèmes à échantillons indépendants***.

Historiquement, le codage de la ***voix/parole*** a été particulièrement éclairant car à la base on a un modèle physique/physiologique qui permet d'établir ***un modèle paramétré***, qui permet d'aborder le codage/réduction d'information à partir de ces paramètres. Ainsi, on se trouve plus naturellement dans un cadre de Fisher de construction de modèles. Mais, il y a une vision plus générale du problème, si on l'aborde du coté de ***l'audio***, qui consiste à capturer n'importe quel type de son. Et là, on n'a plus vraiment de modèle *a priori* sur lequel se baser, car d'une part la source est de toute nature, et d'autre part la propagation du signal rend encore plus ***difficile la modélisation*** et amène à concevoir une autre méthodologie, en particulier celle des représentations dans des ***bases orthogonales*** qui vise à ***décorréler les coefficients du signal dans ces représentations***. Ainsi, a été mis en œuvre la notion de ***compression par transformée orthogonale***. Ce point de vue, à l'avantage de fonctionner pour n'importe quel type de signal, dont les images avec les standards JPEG/JPEG2000. Ces standards se distinguent essentiellement par le choix de la base de représentation: JPEG c'est la DCT (Discret Cosine Transform) qui est utilisé, pour JPEG2000 il s'agit d'une base d'Ondelettes [79].

Si l'on prend quelques exemples, considérant la parole dans le cadre de la téléphonie avec une qualité suffisante pour reconnaitre et comprendre l'interlocuteur, le spectre de Fourier considéré est typiquement restreint à la plage $[200, 3400]$ Hz, alors que l'oreille humaine (jeune) entend dans la plage $[20, 20k]$ Hz. Donc, on restreint les très basses fréquences, et on se limite dans les hautes fréquences aux 2 premières harmoniques ce qui permet de reconnaitre toutes les voyelles. Mettons alors que la fréquence maximale soit de 4 kHz, on échantillonne le signal à raison de 8,000 échantillons par seconde (Th.

79. Voir Cours de 2018, 2020 et 2021.

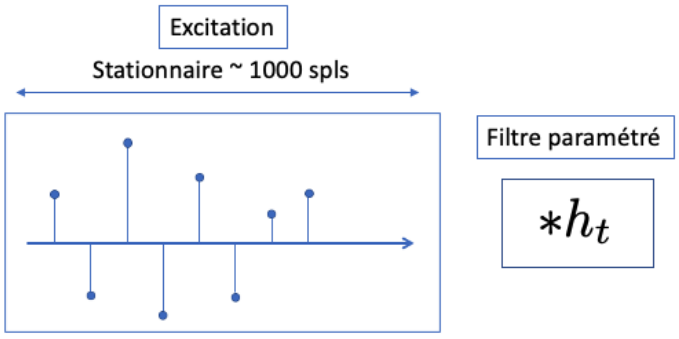

FIGURE 21 – Traitement typique pour la voix dans le cadre téléphonique où l'on considère des filtres paramétrés h_t stationnaires sur une échelle de 1,000 échantillons.

de Nyquist-Shannon [80]), et si chaque échantillon est codé sur un mot de 8 bits, alors on aurait un flux de 64 kb/s (kbits/sec). Il s'agit là du débit qu'il faudrait si l'on faisait ce type de codage assez brutal. Cependant, pour des raisons de coût économique, on va vouloir baisser ce débit tout en gardant la qualité de transmission. Actuellement, pour de la voix sur IP, on arrive à obtenir une réduction significative et il suffit d'un flux d'environ 2.5 kb/s. Pour ce faire, on utilise des modélisations de l'excitation et de la réponse des différents processus physiques, considérés comme des filtres stationnaires sur des échelles de 1,000 échantillons (Fig. 21) afin de mettre en place le codage.

Si l'on considère de l'audio qui est un cadre plus général que celui de la voix sur IP, par exemple considérons la transmission de morceaux de musique, alors on veut des signaux de bien meilleure qualité (haute fidélité). Les CD-Audio qui ont définit les standards dans les années 1980 couvraient une gamme de $[0, 20k]$ Hz soit la totalité de la sensibilité de l'oreille humaine. Donc, l'échantillonnage était de 44.1 kHz et chaque échantillon était codé sur 16 bits, d'où un débit de 706 kb/s. Cependant, pour diffuser de la musique en temps réel (streaming), il nous faut comprimer l'information sans dégrader la qualité d'écoute. L'idée cette fois est d'utiliser des bases orthogonales qui restaurent la haute fidélité avec des débits de l'ordre de 100 kb/s. Si on veut diminuer encre plus le débit, on affecte alors la qualité de restitution.

Concernant, l'imagerie (statique) dont la taille est typiquement 1024x1024 pixels

80. Cours 2021 Sec. 6.4

dont chacun est codé par exemple sur 8 bits (1 byte), cela donne typiquement 1 MB de données. Or, pour gagner de la place sur les supports d'archivage, la compression JPEG descend à 0.5 bit/pixel, soit un gain entre 10 à 20 sans grosse dégradation. Le gain du passage à JPEG2000 est notable quand on veut obtenir de forts taux de compression.

Enfin pour la vidéo, on pourrait penser le flux comme une succession d'images 2D, en faisant une simple extension 3D par regroupement de toutes les images en 1 seul bloc. Cette façon de voir la chose est assez inefficace, car en général lors d'une vidéo il y a une scène fixe (invariante) dans laquelle il y a quelques éléments qui bougent, donc il y a une différence notable entre la variable temps et les 2 variables d'espace. La prise en compte de ces spécificités a été la base du codage MPEG (1988). Schématiquement, on tente de calculer le champ de vitesses (flot optique) de déplacements des pixels d'une image à la suivante, puis on code ce champ. Ainsi, à partir d'une image on peut prédire ce que sera la scène, puis on effectue la soustraction avec l'image réelle pour obtenir une image d'erreur, laquelle est codée comme dans le cas de l'image statique, le plus souvent en JPEG. Ce qui demande le plus de bits, c'est de coder l'image d'erreur, car le champ de vitesse est finalement assez léger, car en principe peu de chose bouge.

8.2 La distorsion et hypothèse de haute résolution

A la base de la prise de sons, images, etc, il y a une ***numérisation*** qui est faite qui nous fait passer de valeurs réelles à nombre infini d'information, à des entiers sur 8, 16... bits. Il nous faut donc traiter la distorsion que cette numérisation implique. Commençons par une *v.a* X car cela va nous permettre de faire le lien entre l'entropie sur des alphabets de taille finie et l'entropie différentielle qui s'applique à des *v.a* réelles. Donc, soit $p(x)$ la densité de probabilité de X, et la numérisation revient à définir un ***quantificateur*** Q qui consiste à découper l'axe réel en boite possiblement de taille variable (Fig. 22):

$$Q(x) = a_k \quad \text{si } x \in]y_{k-1}, y_k] \tag{239}$$

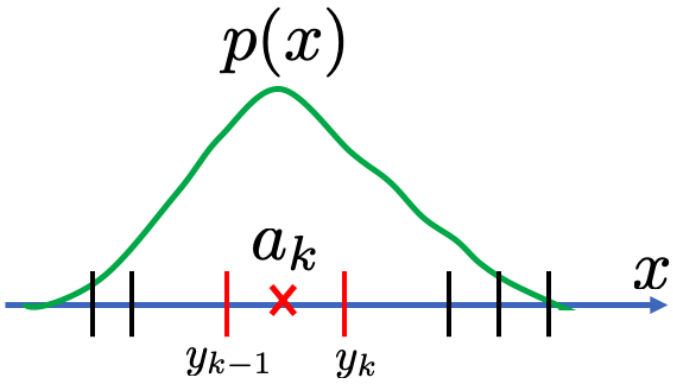

FIGURE 22 – Schématisation de l'opérateur de quantification (Eq. 239).

Bien entendu cet opérateur (non linéaire) va introduire une erreur, et l'on définit la distorsion [81] (nb. norme quadratique) selon l'expression

$$\boxed{D = \mathbb{E}(\|X - Q(X)\|^2) = \int (x - Q(x))^2 p(x) dx} \tag{240}$$

Bien entendu, on a envie de minimiser cette distorsion, ce qui revient à se demander comment choisir les boites $(y_k)_k$. Pour ce faire, nous allons faire une hypothèse de régularité simplificatrice dans un premier temps, avant d'aborder ses limites. Il s'agit de ***la quantification haute résolution*** consistant:

$$\forall x \in]y_{k-1}, y_k], \quad p(x) \approx p(a_k) \qquad \text{(Hte Résol.)} \tag{241}$$

Ainsi,

$$p_k = \mathbb{P}\{Q(X) = a_k\} = \mathbb{P}\{x \in]y_{k-1}, y_k]\} = \int_{y_{k-1}}^{y_k} p(x) dx \approx p(a_k)(y_k - y_{k-1}) = p(a_k)\Delta_k \tag{242}$$

81. En soit, la norme quadratique n'est pas un mauvaise mesure, mais dans un certains cas il y en a de meilleures. Voir discussion Sec. 8.6.

Théorème 20

Sous hypothèse de quantification haute résolution alors $a_k = (y_k + y_{k-1})/2$ *et la distorsion s'écrit*

$$\boxed{D = \sum_k p_k \frac{\Delta_k^2}{12}} \tag{243}$$

La démonstration procède par la transcription de l'expression de D selon l'hypothèse en question:

$$\begin{aligned} D &= \int (x - Q(x))^2 p(x) dx \\ &= \sum_k \int_{y_{k-1}}^{y_k} (x - a_k)^2 p(a_k) dx \overset{min}{=} \sum_k p(a_k) \frac{2}{3} \frac{\Delta_k^3}{2^3} = \sum_k p_k \frac{\Delta_k^2}{12} \end{aligned} \tag{244}$$

(Nb. il n'est pas étonnant que l'on retrouve la contribution des variances des distributions uniformes sur chaque boîte de largeur Δ_k.). Dans le cas d'un quantificateur uniforme où tous les Δ_k sont constants pris égaux à Δ, alors

$$\boxed{D = \frac{\Delta^2}{12} \qquad \text{(Quantif. constant)}} \tag{245}$$

L'erreur est indépendante de la distribution $p(x)$ sous-jacente mais *à condition que l'on puisse opérer l'hypothèse de haute résolution.* Nous y reviendrons.

8.3 Quantificateur optimal

A présent, le problème de la compression se pose en ces termes: ***comment choisir le bon quantificateur soit pour minimiser le nombre de bits pour une erreur fixée, soit pour minimiser l'erreur étant donné le nombre de bits fixé***. Prenons cette seconde version du problème. On pourrait se dire que l'on va diminuer la taille des boites là où la probabilité est importante, et vice versa là où la probabilité est faible, on aurait tendance à vouloir agrandir la taille de la boite. Est-ce la bonne réponse? *Attention, on se fixe le*

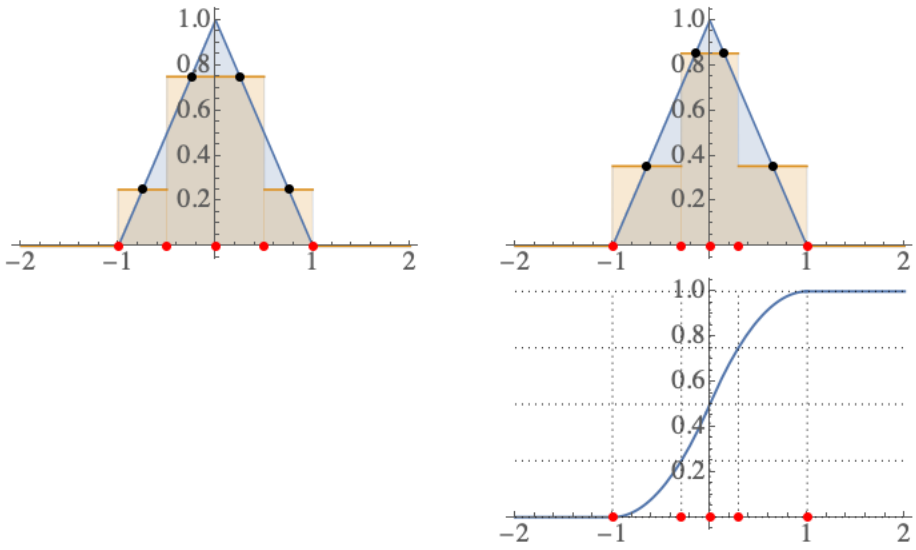

FIGURE 23 – Exemple de deux stratégies découpage de l'axe des x de densité de probabilité $p(x)$ ici triangulaire sur $[-1, 1]$: à gauche, il s'agit du découpage uniforme ou quantification constante, à droite il s'agit du découpage effectué pour obtenir une probabilité constante dans chaque boite. Ce dernier est obtenu en considérant la fonction de répartition $F(x)$ (en bas à droite) et un découpage uniforme de l'axe des y.

nombre de bits! Or, cette contrainte n'est pas équivalente à celle de se fixer le nombre de boites de quantification (dont la solution optimale aurait été celle avancée).

Si on fait un codage "naif", et que l'on se fixe le nombre de boites (K) alors le nombre de bits est de l'ordre de $\log_2 K$. Mais, on sait que l'on peut faire mieux car à un epsilon près le nombre de bits moyen du codage optimal est donné par l'entropie $\mathbb{H} = -\sum_k p_k \log_2 p_k$. Ce que l'on va constater c'est que ***le quantificateur optimal est le quantificateur constant***. Pourquoi? La raison tient au petit raisonnement du début de la section 6.2. Prendre des boites de tailles variables tel que décrit ci-dessus, implicitement optimise le fait que la probabilité de chaque boite est uniforme ce qui induit une grande entropie, par contre prendre un quantificateur uniforme reflète la probabilité $p(x)$ dans chaque boite, et l'on diminue drastiquement l'entropie (à la limite si $p(x)$ est dans 1 seul bin, l'entropie est nulle). On pourrait arguer que l'on perd par contre en erreur avec la solution du quantificateur uniforme. Donc, il faut procéder au calcul...

NDJE: Pour illustrer numériquement le propos, on peut se reporter à l'exemple de la figure 23. Dans le cas de la quantification uniforme (graphe en haut à gauche), l'entropie et la distorsion valent respectivement $\mathbb{H} = 3/4(1 + \log_2(8/3)) \approx 1.81$ *et* $D = 1/48 = 0.0208$*, tandis que pour le schéma de quantification de probabilité uniforme (graphes de droite), alors* $\mathbb{H} = 2$ *et* $D = (2 - \sqrt{2})/24 = 0.0244$*. Les valeurs de* $\mathbb{H}$ *et* D *bien qu'approximativement les mêmes pour les deux schémas de quantification, néanmoins la quantification uniforme est la meilleure. Notons que l'entropie différentielle dans ce contexte vaut* $2/\log 16 \approx 0.72$*.*

Théorème 21 *(**Quantificateur optimal**)*
Soit $p(x)$ *la densité probabilité de* X*, d'entropie différentielle*

$$\mathbb{H}_d[X] = -\int p(x) \log_2 p(x) dx \tag{246}$$

Sous hypothèse de quantification haute résolution, on peut définir l'entropie dite de la source $Q(X)$ *qui est aussi le nombre de bits* R *qu'il faut pour la coder*

$$\mathbb{H}[Q(X)] = -\sum_k p_k \log_2 p_k = R \tag{247}$$

alors

$$\mathbb{H}[Q(X)] \geq \mathbb{H}_d[X] - \frac{1}{2} \log_2(12D) \tag{248}$$

et on a égalité ssi la quantification est uniforme de pas Δ*, elle minimise* $\mathbb{H}$ *et l'on a*

$$\boxed{R = \mathbb{H}[Q(X)] = \mathbb{H}_d[X] - \log_2 \Delta} \tag{249}$$

Ainsi, le quantificateur constant est le quantificateur optimal.

Par le biais de ce théorème, on obtient le lien entre l'entropie différentielle (cadre continu) et l'entropie (cadre discret). Si on se souvient que le volume d'un ensemble typique est environ $V \approx 2^{nH_d(x)}$, et que le volume en dimension n de la boite autour d'un élément de cet ensemble $\delta V \approx \Delta^n$ alors la relation ci-dessus nous donne en fait $\log_2 N_b = n\mathbb{H}$ où N_d est le nombre de boites de quantification qu'il faut pour paver tout l'espace typique. Passons à la démonstration, mais avant il valait mieux se faire une idée du résultat pour bien le comprendre.

Démonstration 21. Donc, on veut calculer $\mathbb{H}$, c'est-à-dire

$$\mathbb{H} = -\sum_k p_k \log_2 p_k \tag{250}$$

sous l'hypothèse de haute résolution, donc $p_k = p(a_k)\Delta_k$, et sur l'intervalle $]y_{k-1}, y_k]$ la probabilité est constante $p(x) = p(a_k)$. Ainsi,

$$\begin{aligned}
\mathbb{H} &= -\sum_k \log_2(p(a_k)\Delta_k) \times \int_{y_{k-1}}^{y_k} p(x)dx \\
&= -\sum_k \int_{y_{k-1}}^{y_k} \log_2(p(x)\Delta_k)p(x)dx \\
&= -\sum_k \int_{y_{k-1}}^{y_k} p(x)\log_2 p(x)dx - \sum_k \int_{y_{k-1}}^{y_k} \log_2(\Delta_k)p(x)dx \\
&= \mathbb{H}_d - \sum_k p_k \log_2(\Delta_k)
\end{aligned} \tag{251}$$

Etant donné que la densité de probabilité est fixée, $\mathbb{H}$ et $\mathbb{H}_d$ sont fixées, et seules les valeurs des $(\Delta_k)_k$ et les p_k restent à optimiser sachant que $\sum_k p_k = 1$. Donc, nous avons une minimisation sous contrainte. Mais la fonction à minimiser est la moyenne de log, c'est-à-dire $-\sum_k p_k \frac{1}{2}\log_2(\Delta_k^2)$. Or, $-\log(x)$ étant une fonction strictement convexe en utilisant l'inégalité de Jensen (Th. 12), il vient

$$\mathbb{H} \geq \mathbb{H}_d - \frac{1}{2}\log_2\left(\sum_k p_k \Delta_k^2\right) = \mathbb{H}_d - \frac{1}{2}\log_2(12D) \tag{252}$$

Ce qui donne le résultat escompté, mais on conviendra que si nous avions uniquement procédé à la démonstration, sa signification n'aurait pas été révélée au premier abord. De plus, l'inégalité de Jensen devient une égalité ssi $\Delta_k = Cte = \Delta$. ce qui fournit le 2nd résultat. **L'optimum est réalisé pour le quantificateur constant.** ■

NDJE: Notons que dans le cas (quantificateur constant) de l'exemple de la figure 23, on a $\mathbb{H}_d = 2/\log(16) \approx 0.72$ *et donc* $\mathbb{H}_d - 1/2\log_2(12D) = 1.72$ *alors que* $\mathbb{H} \approx 1.81$. *Si on augmente le nombre de boites de quantification, de 4 à 10, alors* $\mathbb{H}_d - 1/2\log_2(12D) \approx 3.04$ *tandis que* $\mathbb{H} = 3.06$ *(accord relatif à* $7\ 10^{-3}$*), et pour* 100 *boites l'accord relatif est bon à* $5\ 10^{-5}$ *prés. Ainsi, l'accord devient parfait asymptotiquement.*

Si on revient au problème de codage, si la distorsion (erreur) D est fixée alors le nombre de bits moyen est égal à

$$R(D) = -\sum_k p_k \log_2 p_k = \mathbb{H} = \mathbb{H}_d - \frac{1}{2}\log_2(12D) \tag{253}$$

ou bien si R est fixé, on obtient une erreur minimale donnée par

$$\boxed{D(R) = \frac{1}{12} 2^{2(\mathbb{H}_d - R)}} \tag{254}$$

Ce qui donne ***une décroissance exponentielle*** de D en fonction de R: ex. si on ajoute 1 bit, l'erreur quadratique est divisée par un facteur 4. Ainsi, ce résultat nous donne le résultat de l'erreur lorsque que l'on a ***1 v.a réelle codée avec le quantificateur optimal***. C'est le résultat de base. Cependant, le problème qui se présente en pratique est celui d'un signal, possiblement en grande dimension, qui a de ***la redondance et de la structuration***, c'est-à-dire deux caractéristiques que l'on aimerait exploiter. Nous allons donc essayer de trouver une représentation la mieux adaptée à ces signaux. La technique la plus simple et algorithmiquement efficace est l'usage de ***bases orthonormales***. La thématique a fait l'objet du Cours en 2021 mettant en relation *Parcimonie, Régularité et Approximation*. Cette fois-ci nous allons prendre le biais du ***codage par transformée orthogonale***. Rappelons que ce codage est très bien adapté pour des signaux dont la structuration n'est pas suffisante pour tenter des modèles paramétrés dédiés. Avant cela il faut procéder à la quantification du signal.

8.4 Quantification scalaire

Soit Y un vecteur de taille N $(0 \leq n < N)$ dont les composantes sont notées $Y[n]$, et soit une base orthonormale dans $\mathbb{R}^N$, $\mathscr{B} = \{g_m\}_{0 \leq m < N}$ avec

$$\langle g_m, g_{m'} \rangle = \sum_n g_m[n]\ g^*_{m'}[n] = \delta(m - m') \tag{255}$$

La décomposition du vecteur Y se lit

$$Y = \sum_m \langle Y, g_m \rangle \; g_m = \sum_{m=0}^{N-1} A[m] \; g_m \tag{256}$$

Ainsi, au lieu de coder les composantes de Y dans une base canonique de Dirac, ***on se donne la liberté de choisir une base*** $\mathscr{B}$ ***et de coder les coefficients de*** Y ***dans celle-ci*** (c'est-à-dire le codage des produits scalaires). Attention, au passage les $A[m]$ sont des variables aléatoires car

$$A[m] = \sum_{n=0}^{N-1} Y[n] \; g_m^*[n] \tag{257}$$

c'est-à-dire que $A[m]$ ***est une combinaison linéaire des*** N ***v.a*** $Y[n]$. Dans la suite les majuscules X, Y sont des v.a tandis que les minuscules sont des scalaires.

Donc, le problème du codage du vecteur Y revient à résoudre celui du codage des $A[m]$. On commence par supposer que

$$\mathbb{E}[A[m]] = 0 \tag{258}$$

Si tel n'est pas le cas, on suppose que l'on peut soustraire ces moyennes et que récepteur et émetteur de Y peuvent les stocker une fois pour toute.

L'étape suivante est de procéder à la ***quantification scalaire***, où on va opérer sur 1 seule composante à la fois, c'est-à-dire

$$\boxed{\hat{A}[m] := Q(A[m])} \tag{259}$$

On aurait pu considérer le block des N valeurs $A[m]$ et tenter d'adapter en dimension N la boite de quantification [82]. En effet, en dimension N rien n'empêche de prendre des pavages non réguliers adaptés à la distribution de probabilité. Le cas de quantification scalaire est beaucoup plus simple, cela revient à prendre des pavés en dimension N. On pourrait imaginer que cette simplification est brutale, et qu'il y a mieux à faire. Ce que nous dit S. Mallat est qu'après beaucoup de travaux de recherche dans ce domaine, en grande dimension procéder à une optimisation du pavage n'apporte finalement pas grand chose. En fait, ***le gain de ces optimisations de pavage, est compensé par l'optimisation***

82. c'est la quantification vectorielle.

de la représentation, plus complexe que les bases orthonormales, tout en se fixant une quantification scalaire. Jusqu'à 3-4 ans le problème semblait figé avec des standards en audio/imagerie etc, mais depuis il y a de nouveaux développements car les réseaux de neurones arrivent à mieux comprimer les signaux, donc ils ont capturer quelque chose qui manquait aux représentations antérieures. Cependant, même dans ce (nouveau) contexte, la quantification est néanmoins scalaire.

Cette quantification introduit une erreur comme on a vu à la section précédente et le vecteur $\hat{Y}$ construit à partir des $\hat{A}[m]$ (signal reçu au mieux) s'écrit

$$\hat{Y} = \sum_m \hat{A}[m] g_m \tag{260}$$

Delà, l'erreur quadratique entre Y et $\hat{Y}$ est facilement calculable grâce à l'orthonormalité de la base $\mathscr{B}$

$$D = \mathbb{E}[\|Y - \hat{Y}\|^2] = \mathbb{E}[\sum_m \|A[m] - \hat{A}[m]\|^2] = \sum_m \underbrace{\mathbb{E}[\|A[m] - \hat{A}[m]\|^2]}_{D_m} \tag{261}$$

C'est-à-dire que la distorsion totale est la somme des distorsions de quantification sur chacune des composantes $A[m]$ de Y dans la base $\mathscr{B}$. Le nombre total de bits nécessaires pour coder Y est la somme du nombre de bits nécessaires pour coder chacune des composantes quantifiée, noté R_m. Donc, finalement

$$D = \sum_m D_m \quad \text{et} \quad R = \sum_m R_m \tag{262}$$

Ainsi, on retrouve la problématique développée pour 1 *v.a*, cette fois dans le cadre élargi à N *v.a*, à savoir ***quelle est la distorsion*** D ***totale considérant un nombre total de bits*** R (et vice versa). C'est ***un problème d'allocation de bits où nous avons les problèmes de codage, de quantification, et in fine de choix de la base***.

8.5 Allocation de bits

On veut donc fixer le nombre de bits total pour coder le vecteur Y (plutôt ses composantes dans ma base choisie), mais en même temps cela nous laisse une certaine

liberté d'optimiser le nombre de bits nécessaires pour chacune des N directions. Ce que l'on sait, c'est que chaque $\hat{A}[m]$ va être codée avec un quantificateur constant pour être optimal, et ce qu'il reste donc à optimiser ce sont les N pas de quantification $(\Delta_m)_{m<N}$. Ainsi, le problème se ramène à se demander s'il faut privilégier certaines directions, pour en diminuer l'erreur? A nouveau l'inégalité de Jensen va nous fournir la réponse, et de nouveau le résultat est très simple et troublant à la fois.

Théorème 22 *(**Allocation optimale**)*
Sous les hypothèses de quantification haute résolution, une distorsion totale D fixée alors le nombre de bits total R est minimum si on fixe tous les pas Δ_m à une unique valeur Δ telle que

$$D = \sum_m D_m = N\frac{\Delta^2}{12} \Leftrightarrow \boxed{\Delta^2 = \frac{12D}{N}} \tag{263}$$

De plus, si le nombre de bits par coefficient est noté $\bar{R} = R/N$, et que l'entropie différentielle moyenne est définie selon

$$\overline{\mathbb{H}}_d := \frac{1}{N}\sum_m \mathbb{H}_d(A[m]) \tag{264}$$

alors la distorsion totale est donnée par l'expression

$$\boxed{D(\bar{R}) = \frac{N}{12}2^{2(\overline{\mathbb{H}}_d - \bar{R})}} \tag{265}$$

Notons que $\overline{\mathbb{H}}_d$ est fixée dès lors que l'on a fixé la base de décomposition et pour mémoire les $A[m]$ (les produits scalaires) sont des *v.a.* Le théorème nous dit alors qu'en N dimension en quantification haute résolution, la quantification constante unique pour toutes les composantes est optimale. On procède à un pavage en petit hypercube d'arête Δ.

Démonstration 22. Le nombre bits est défini par l'entropie (codage entropique), donc pour chaque composante quantifiée $\hat{A}[m]$, il est donnée par le théorème 21. C'est-à-dire

$$R_m = \mathbb{H}(\hat{A}[m]) = \mathbb{H}_d(A[m]) - \log_2 \Delta_m = \mathbb{H}_d(A[m]) - \frac{1}{2}\log_2(12D_m) \tag{266}$$

Si on effectue la moyenne sur les N composantes, on obtient

$$\bar{R} = \overline{\mathbb{H}}_d - \frac{1}{2}\left(\frac{1}{N}\sum_m \log_2(12D_m)\right) \tag{267}$$

Pour optimiser les D_m, de nouveau on se trouve avec une moyenne de log, donc par la stricte convexité de $-\log$, on a

$$\bar{R} \geq \overline{\mathbb{H}}_d - \frac{1}{2}\log_2\left(\frac{12}{N}\sum_m D_m\right) = \overline{\mathbb{H}}_d - \frac{1}{2}\log_2\left(\frac{12D}{N}\right) \tag{268}$$

Il y a égalité, et donc optimisation puisqu'on atteint alors la borne inf., si $D_m = cte = D/N$ ce qui a pour conséquence que tous les Δ_m sont tous égaux à un Δ tel que

$$\Delta^2 = \frac{12D}{N} \tag{269}$$

ce qui est le premier résultat. De même en écrivant l'égalité, on obtient

$$\bar{R} = \overline{\mathbb{H}}_d - \frac{1}{2}\log_2\left(\frac{12D}{N}\right) \tag{270}$$

ce qui fournit le second résultat du théorème. ■

Donc, ce théorème nous dit que ***dans le cadre de quantification haute résolution***, la solution optimale est la plus simple, c'est-à-dire celle obtenue avec un pas unique obtenu soit à D fixée soit à R fixé. ***On a un pavage régulier en hypercubes selon des axes de la base orthonormale.*** Donc, la question qui se pose maintenant est celui du ***choix de la base***.

Comment va-t'on procéder? Quand on regarde le théorème ci-dessus, on remarque que le seul lien avec la base est le facteur $\overline{\mathbb{H}}_d$, c'est-à-dire l'entropie différentielle des produits scalaires de Y avec les vecteurs de base. Or, $\overline{\mathbb{H}}_d$ est la moyenne des entropies différentielles des distributions de probabilité des composantes (non quantifiée) $A[m]$ ($p_m(x)$). On sait que l'entropie est d'autant plus petite que la probabilité est concentrée autour de sa valeur moyenne [83]. Or, par construction la moyenne est nulle ($\int p_m(x)dx = 0$), donc la concentration attendue est autour de la valeur 0. Or, ***c'est précisément le cas qui se***

83. Rappelons nous du petit exemple de la section 6.2.

produit quand on a une forme de sparsité engendrée par une décomposition parcimonieuse. Notons que comme il y a une conservation d'énergie dans une base orthonormale, s'il y a beaucoup de 0, en contrepartie il y a de grands coefficients par ailleurs (mais ils sont peu nombreux). Le hic que nous allons voir, c'est qu'en poussant le raisonnement de parcimonie jusqu'au bout, cela va entrer en conflit avec l'hypothèse de haute résolution, ce qui va nous amener à revisiter les résultats obtenus jusqu'alors.

Mais avant de pousser le bouchon un peu trop loin, exploitons déjà les résultats obtenus afin d'optimiser la base orthonormale, car ils ont été à la base de toutes les idées de codage jusqu'aux années 90s.

8.6 Choix de la base orthonormale

Si on se rappelle de l'expression de R_m (Eq. 266), il faut faire attention à la présence de l'entropie différentielle car celle-ci peut être négative (contrairement à son homologue non différentielle). Or, on n'a ni imposé que R_m soit positif, ni même qu'il soit un entier. Donc, en termes de nombre de bits, il va falloir être précautionneux. En quelque sorte, il nous faut trouver des équivalents au codage de Huffman (Sec. 7.3) qui satisfont $R \in \mathbb{N}$. Par exemple, on va utiliser des algorithmes "gloutons" (*greedy*)[84] (*NDJE: voir un exemple Sec. 8.7*) qui allouent successivement un bit parmi les différentes composantes en minimisant la distorsion globale. On trouve la solution optimale, car le problème est convexe. Ce que l'on constate par ailleurs c'est un comportement asymptotique (N grand) comparable à celui du théorème 22.

Un autre aspect avant d'aborder le choix de base, consiste à remarquer que pour certaines applications tenir compte des erreurs pour certaines composantes n'a pas lieu d'être: ex. si on abouti dans le cas de l'audio, à des codages qui génèrent de grandes erreurs pour des fréquences au-delà de 20 kHz, à la limite ce n'est pas crucial car l'oreille humaine sera bien incapable de les détecter. Donc, la perception du dispositif de réception est à prendre en compte. Traduit en termes plus mathématique, cela revient à se poser

84. D'une manière générale un algorithme "glouton" est un algorithme qui réalise un choix optimal à chaque étape sans se préoccuper ni de ce qui a précédé, ni de ce qui adviendra par la suite, en espérant obtenir le résultat optimal global mais il n'y a pas de garantie. Le code de Huffman en est un exemple. D'autres exemples en dehors du codage: rendu de monnaie, organiser au mieux l'occupation de salles, obtenir le parcours du voyageur de commerce, déterminer le plus court chemin dans un réseau...

la question de savoir si la norme $L2$ est adaptée pour le cas qui nous concerne. La façon dont ***on peut tenir compte de la perception est d'utiliser des normes pondérées***. Donc, on remplace l'expression de D (Eq. 262) par

$$D_w = \sum_m \frac{D_m}{w_m^2} = \sum_m D_m^w \tag{271}$$

avec des poids $1/w_m^2$. Au passage, si on s'imagine m comme un axe fréquentiel (ou équivalent) alors w_m tient compte de la réponse en fréquence du récepteur.

Maintenant, comment les résultats précédents se transforment-ils si on optimise D_w? Remarquons que

$$D_m^w = \mathbb{E}\left[\frac{1}{w_m^2}\|A[m] - \hat{A}[m]\|^2\right] = \mathbb{E}\left[\left\|\frac{A[m]}{w_m} - \frac{\hat{A}[m]}{w_m}\right\|^2\right] \tag{272}$$

Cela revient donc à pondérer les coefficients par les w_m, mais attention $Q(w_m^{-1}A[m]) \neq w_m^{-1}Q(A[m])$ en toute généralité. Cependant, ce qui est optimal d'après ce que l'on a vu dans les théorèmes précédents, c'est de procéder à une *quantification constante* (uniforme) des $A[m]w_m^{-1}$ avec un pas Δ qui est équivalent [85] à quantifier $A[m]$ avec un pas

$$\Delta_m = w_m\Delta \tag{273}$$

Ainsi, on met au point des ***quantifications "non uniformes" adaptée au problème particulier***, non pas à cause de la stratégie optimale d'allocation de bits dans le cas d'une erreur en norme $L2$, mais parce qu'au contraire ***on adapte la métrique en pondérant les distorsions, ce qui permet d'allouer plus d'erreur à certains canaux***. Plus $1/w_m^2$ est grand (petite valeur de w_m comme une meilleure définition de perception du signal), plus la contribution à D est importante, il nous faut alors accorder un plus petit pas de quantification. A l'inverse, de grandes valeurs de w_m (moins bonne définition), $1/w_m^2$ est petit donc une faible contribution à l'erreur totale, et l'on peut quantifier avec un grand pas. Nous reviendrons sur ce point, car on a des cas où la physiologie neurologique nous indique qu'il parait judicieux d'adapter l'erreur allouée pour chaque canal (fréquence) en fonction du signal lui-même.

85. C'est tout simplement dire que $(A[m]w_m^{-1})/\Delta = A[m]/(w_m\Delta)$.

Pour choisir la base orthonormale, nous allons exploiter ***la redondance du signal***. On suppose que le signal est ***régulier par morceaux***, et supposons que cela soit le temps que l'on découpe. On veut trouver une base orthonormale dans laquelle l'entropie différentielle du signal soit la plus petite possible. Or qui dit signal très régulier, signifie que la décroissance des coefficients de Fourier à haute fréquence est rapide [86]. De la décroissance du spectre, on peut lire la classe de régularité de la fonction (au sens de Sobolev). Ainsi, sur chaque intervalle (de temps), on procède à une discrétisation en N points, et la base orthonormale de Fourier discrète sur $\mathbb{R}^N$ est donnée par l'ensemble des vecteurs $\{g_k\}_{k<N}$ suivants [87]

$$\boxed{\mathscr{B}_F = \left\{ g_k(n) = \frac{\exp\left\{ i\frac{2\pi k}{N} n \right\}}{\sqrt{N}} \right\}, \qquad (k,n) \in [\![0, N-1]\!]} \tag{274}$$

Mais est-ce que cela va marcher? ou plutôt est-on certain qu'il n'y aura pas de hautes fréquences? La remarque à faire c'est qu'implicitement, la définition des $g_k(n)$ induit une condition de périodicité sur $\mathbb{Z}$ en dehors de $n = 0, \ldots, N-1$. Or, le signal dans l'intervalle en question n'a pas lieu d'être périodique, et il y a donc ***une discontinuité aux bords***, celle-ci entraine alors un spectre en $1/\omega$ ou en $1/k$ ($\omega_k = 2\pi k$). Ce qui a pour conséquence de dépenser beaucoup de bits à coder la discontinuité. Il nous faudra donc opter pour ***la base de cosinus***.

8.7 NDJE: exemple d'algorithme glouton d'allocation de bits

Je vais particulariser le théorème 22 dans le cas où les $A[m]$ sont des $v.a$ indépendantes avec chacune une distribution gaussienne $\mathcal{N}(0, \sigma_m^2)$. Dans ces conditions, l'entropie différentielle de chaque composante est égale à

$$\mathbb{H}_d(m) = \frac{1}{2} \log_2(2\pi e \sigma_m^2) \tag{275}$$

86. Voir par ex. Cours 2021 Sec. 3.3, Cours 2019 Sec. 5.3.1, Cours 2018 Sec. 5.2.3.

87. NDJE: pour se rapprocher des notations des sections antérieurs, il faut plutôt voir des $(g_m)_m$ en lieu et place des $(g_k)_k$.

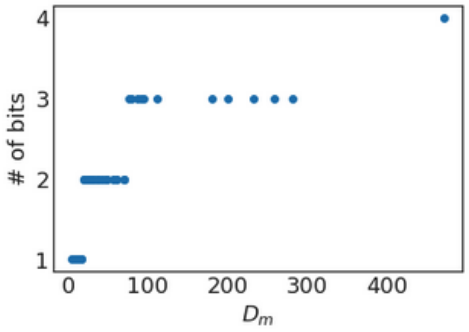

FIGURE 24 – Résultat de l'algorithme 1 d'allocation de bit pour $N = 50$ et $R = 100$.

Et la distorsion, D_m associée à la quantification optimale est reliée au nombre de bits alloués R_m via

$$D_m = \frac{1}{12} 2^{2(\mathbb{H}_d(m) - R_m)} = \frac{\pi e}{6} \sigma_m^2 2^{-2R_m} \tag{276}$$

Remarquons que si $R_m = 0$ alors $D_m \propto \sigma_m^2$ ($c = \pi e/6$) étant la constante de proportionnalité), et ajouter un bit la diminue d'un facteur 4. Donc, le problème est connaissant les valeurs $(\sigma_m^2)_m$ comment allouer les R_m bits à chaque composante m, sachant que $\sum_m R_m = R$ est fixé. Un algorithme "glouton" simple [88] est de procéder par itération à l'allocation de un bit supplémentaire à la composante ayant la plus grande distorsion afin de la diminuer d'un facteur 4:

Algorithme 1 *(**Allocation de bits**)*

initialisation *: $\forall m = 0, \dots, N-1$, $D_m = \sigma_m^2$, $R_m = 0$;*
while-loop *: sous la condition $\sum_{m=0}^{N-1} R_m < R$, procéder successivement à :*

- $m = \underset{n}{\text{argmax}}\ D_n$
- $R_m \leftarrow R_m + 1$
- $D_m \leftarrow D_m/4$

Un exemple du résultat de cet algorithme est donné sur la figure 24. Ensuite, on peut comparer la valeur de la distorsion totale obtenue en sommant tous les D_m, à l'expression

88. Voir notebook `Allocation_de_bits.ipynb`.

de D optimisée:

$$D^{algo} = \sum_{m=0}^{N-1} D_m \qquad\qquad D^{optim} = N\frac{\pi e}{6}\left(\prod_{m=0}^{N-1}\sigma_m^2\right)^{1/N} 4^{-R/N} \tag{277}$$

Dans le cas de l'exemple, on obtient pour une configuration des σ_m^2, les valeurs suivantes de distorsions, $D^{algo} \approx 88.25$ et $D^{optim} \approx 81.08$, ce qui donne une efficacité de 91.9%.

9. Séance du 9 Mars

9.1 Rappels de la séance précédente

Nous allons voir des applications du corpus théorique élaboré dans le cours de cette année. Et nous allons trouver des résultats inattendus qui remettent en question ce corpus. Afin de comprendre l'origine des différences, nous allons revisiter la parcimonie qui est au cœur du problème. Par ce biais, nous ferons le lien avec les cours des années précédentes en particulier celui de 2021 sur les représentations.

La séance dernière nous avons motivé l'usage de ***la compression par transformée orthogonale*** quand on ne dispose pas de modèle sous-jacent à la production des observations/données (ex. l'audio qui concerne tous les types de sons, en contre-point de la parole que l'on peut modéliser). Ainsi, si l'on dispose d'un vecteur Y à N composantes $Y[n]$ $(0 \leq n < N)$, la représentation dans une base orthonormale est donnée par les produits scalaires de Y avec les vecteurs unitaires de la base $\mathscr{B} = \{g_m\}_{m<N}$, noté $\langle Y, g_m\rangle = A[m]$. Nous avons également vu que ***la quantification constante (uniforme) de ces composantes est optimale***, et nous avons pu mettre au point un petit algorithme *glouton* d'allocation presque-optimale de bits.

Le signal reconstruit à partir des composantes quantifiées n'est pas identique au signal d'origine

$$\hat{Y} = \sum_m \hat{A}[m] g_m \tag{278}$$

et pour optimiser la quantification, nous avons utilisé comme mesure de ***la distorsion*** une erreur quadratique qui n'est autre que la somme des distorsions sur chacune des

composantes:

$$D = \mathbb{E}[\|Y - \hat{Y}\|^2] = \sum_m \mathbb{E}[\|A[m] - \hat{A}[m]\|^2] = \sum_m D_m \tag{279}$$

Et si l'on se fixe le nombre total de bits R pour coder Y, alors $R = \sum_m R_m$ la somme des bits alloués pour chaque composante. L'optimisation nous a fait relier R_m à l'entropie différentielles (Eq. 266) que je redonne ici pour mémoire

$$R_m = \mathbb{H}(\hat{A}[m]) = \mathbb{H}_d(A[m]) - \log_2 \Delta_m = \mathbb{H}_d(A[m]) - \frac{1}{2}\log_2(12D_m) \tag{280}$$

Nous avons vu que tous les pas de quantifications Δ_m sont égaux à Δ (un pas constant pour toutes les composantes), et le théorème 22 nous donne la formule qui relie D au nombre moyen de bits $\bar{R} = R/N$ donnant un scaling en $D \propto 4^{-\bar{R}}$. Rappelons un point imortant que toute cette théorie fait usage d'une ***hypothèse dite de "haute résolution"*** (Eq. 241) qui stipule que pour tous les intervalles de quantification, on peut approximer la densité de probabilité $p_m(x)$ de $A[m]$ par une constante.

Le degré de liberté qui nous reste est relié au pré-facteur de la distorsion D (Th. 22), il s'agit de l'entropie différentielle moyenne

$$\overline{\mathbb{H}}_d := \frac{1}{N}\sum_m \mathbb{H}_d(A[m]) \tag{281}$$

Car en effet, plus la distribution de $A[m]$ est concentrée autour de sa moyenne (prise nulle par construction) plus l'entropie différentielle est petite. Or, ce critère de concentration autour de 0, signifie en creux une sparsité liée à une ***représentation parcimonieuse de*** Y ***dans la base orthonormale*** $\mathscr{B}$. Donc, nous abordons le choix de la base orthonormale.

9.2 Signaux réguliers par morceaux: la DCT

On considère le cas de signaux réguliers par morceaux. Typiquement, on découpe la trame temporelle en intervalles par exemple de taille N sur lesquels on procède au codage du signal. Ce que nous avons abordé à la fin de la séance dernière, c'est que la base de Fourier discrète qui viendrait à l'esprit

$$\mathscr{B}_F = \left\{ g_m(n) = \frac{\exp\left\{i\frac{2\pi m}{N}n\right\}}{\sqrt{N}} \right\}, \qquad (m,n) \in [\![0, N-1]\!] \tag{282}$$

va nous poser problème à cause de la périodicité (N) imposée implicitement par les g_k. Or, le signal lui-même n'a pas lieu de suivre cette périodicité. ***Des discontinuités apparaissent aux bords des intervalles générant alors un spectre avec des composantes à hautes fréquences*** qui sont totalement contre-productives. En effet, non seulement pour un signal régulier, on s'attend à ce que son spectre décroisse rapidement ce qui n'est pas le cas, mais aussi on passe son temps à utiliser des bits pour coder ces discontinuités artificielles. ***Il nous faut donc gommer les discontinuités***.

Pour ce faire, au lieu de forcer la périodisation directe de l'intervalle de signal extrait (N échantillons), on commence par le ***symétriser*** ce qui élimine les discontinuités d'ordre 0. La périodisation de période $2N$ laisse des discontinuités aux bords concernant la dérivée (ordre 1) ce qui est déjà moins gênant (Fig. 25). De plus, la symétrie se fait autour d'un demi-entier, ce qui motive finalement l'emploi de la base

$$\mathscr{B}_{FSym} = \left\{ g_m(n) = \frac{\exp\left\{ i\frac{\pi m}{N}(n+1/2) \right\}}{\sqrt{2N}} \right\}, \qquad (m,n) \in [\![0, N-1]\!] \tag{283}$$

Notant $\tilde{x}(n)$ le signal échantillonné symétrisé

$$\tilde{x}(n) = \begin{cases} x(n) & \text{si } 0 \leq n < N \\ x(-n-1) & \text{si } -N \leq n < -1 \end{cases} \tag{284}$$

il se décompose selon base $\mathscr{B}_{FSym}$, et si l'on sépare la partie réelle et la partie imaginaire

$$\tilde{x}(n) = \sum_{m=0}^{N-1} \alpha_m \cos\left(\frac{\pi m}{N}(n+1/2)\right) + \sum_{m=0}^{N-1} \beta_m \sin\left(\frac{\pi m}{N}(n+1/2)\right) \tag{285}$$

Or, $\tilde{x}(n)$ est paire par rapport à $n = -1/2$, donc la somme sur les sinus est identiquement nulle. Ainsi, la base orthonormale naturelle des signaux réguliers par morceaux est celle des cosinus:

$$\boxed{\mathscr{B}_{cos} = \left\{ g_m[n] = \lambda_m \sqrt{\frac{2}{N}} \cos\left(\frac{\pi m}{N}(n+1/2)\right) \right\}, \qquad (m,n) \in [\![0, N-1]\!]} \tag{286}$$

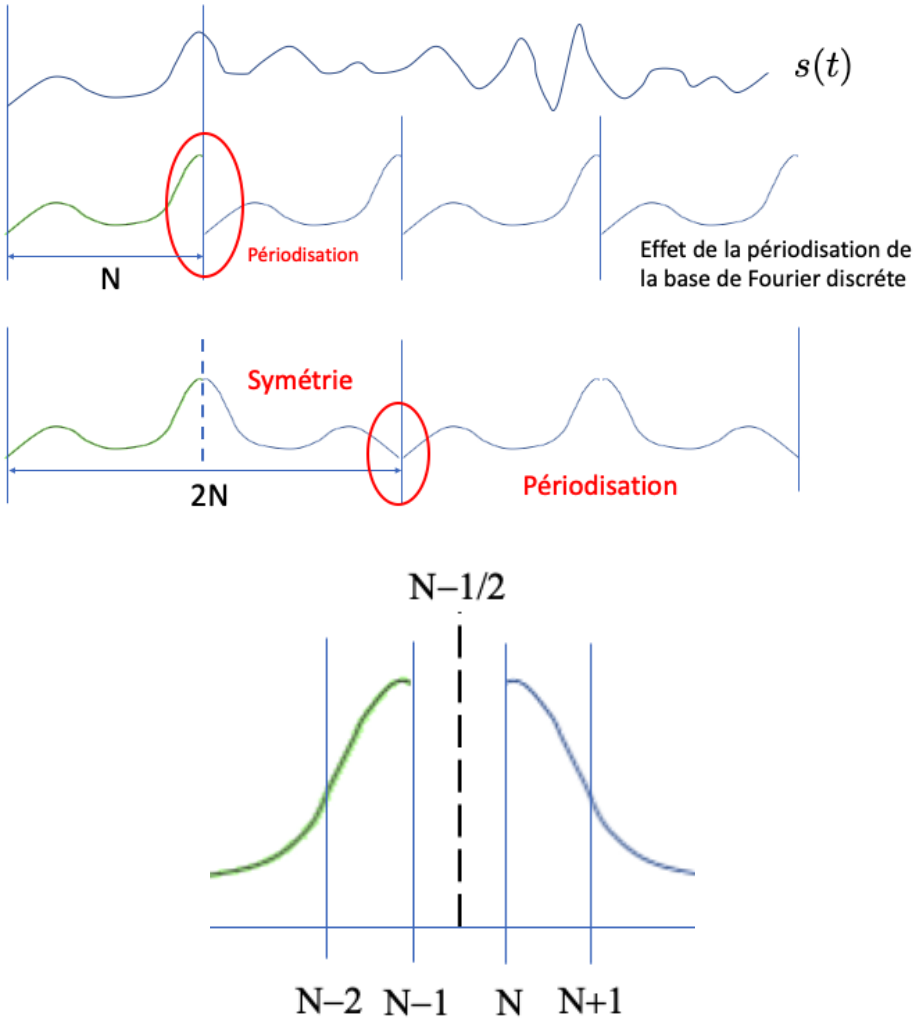

FIGURE 25 – En haut: Schématisation de la procédure de symétrisation pour palier l'effet de périodisation implicite d'une base de Fourier discrète. En périodisant directement le signal extrait, on introduit de grandes discontinuités de jointure. Après la symétrisation, il reste des discontinuités sur la dérivée mais beaucoup moins gênantes. En bas: en zoomant sur le point de jonction de la symétrisation, on s'aperçoit que la symétrie se fait autour d'un demi-entier.

avec le facteur λ_m qui ajuste la normalisation

$$\boxed{\lambda_m = \begin{cases} 1/\sqrt{2} & \text{si } m = 0 \\ 1 & \text{sinon} \end{cases}} \tag{287}$$

Notons que nous avons les relations suivantes:

$$\langle g_m, g_{m'} \rangle = \delta(m - m') = \sum_{n=0}^{N-1} g_m[n] g_{m'}[n] \qquad \sum_{m=0}^{N-1} g_m[n] g_m[n'] = \delta(n - n') \tag{288}$$

Ainsi, la transformée dite ***DCT***, dérivée de la FFT, nécessite $O(N \log_2 N)$ opérations, et est définie selon [89]

$$\boxed{\begin{aligned} x[n] = \sum_{m=0}^{N-1} \check{x}[m] g_m[n] = \sqrt{\frac{2}{N}} \sum_{m=0}^{N-1} \check{x}[m] \lambda_m \cos\Big(\frac{\pi m}{N}(n + 1/2)\Big) \\ \check{x}[m] = \langle x, g_m \rangle = \sqrt{\frac{2}{N}} \lambda_m \sum_{n=0}^{N-1} x[n] \cos\Big(\frac{\pi m}{N}(n + 1/2)\Big) \end{aligned}}$$

9.3 Le cas de l'audio: standard MPEG

Comment appliquée cette transformation au codage audio? La première chose à faire, c'est de découper la trame temporelle en intervalles de taille environ 1024 échantillons. Pour cela on utilise une *fenêtre glissante* de taille fixe [90] $N = 1024$

$$w[n] = 1 \quad \text{si } 0 <= n < N \tag{289}$$

Ainsi, l'extraction des échantillons entre $[\![pN\,, (p+1)N - 1]\!]$ est donnée par la multiplication de $x[n]$ par $w[n - pN]$, puis on applique la DCT ci-dessus. En particulier, cette

89. NDJE: j'ai opté ici pour une forme symétrique, il y a d'autres définitions en pratique et je conseille de se référer à la documentation de chaque librairie.

90. NDJE: j'utilise la notation w pour *window*, cela permet de différentier par rapport aux g_m de la base. D'autre part il peut être judicieux d'utiliser des fenêtres à bords plus doux qu'un rectangle.

enchainement d'opérations fait apparaitre la notion de ***base par blocs***

$$\left\{w[n-pN]\cos\left(\frac{\pi m}{N}(n+1/2)\right)\right\} \qquad \forall m < N, \forall p \in \mathbb{Z} \tag{290}$$

Il y a une extension possible pour laquelle on utilise des tailles de fenêtres différentes, mais l'essentiel est là.

Maintenant, il nous faut ajuster cette base pour obtenir un distorsion minimale. Cependant, l'erreur quadratique n'est pas totalement adaptée au système perceptif, en particulier à cause ***des phénomènes de masquage***[91]. On aimerait ne comptabiliser dans le budget des erreurs uniquement les contributions dont l'origine est un stimuli au-dessus d'un seuil de perception. Mais, le seuil lui-même dépend de la hauteur du stimuli lui-même. En gros, si on stimule l'oreille avec une sinusoïde à fréquence ω, il existe une bande de fréquences autour de ω qui est moins bien perçue, alors qu'en dehors de cette bande l'audition n'est pas affectée. En pratique, en dessous de 700Hz, il existe 7 bandes critique de masquage (de largeur 100Hz), et au-delà de 700Hz les bandes sont de tailles qui s'accroissent (constante en échelle log) au fur et à mesure que la fréquence augmente. L'organe de Corti au centre de la cochlée est couvert de cils baignant dans un liquide, et la réponse de ces cils se modélise comme ***des filtres passe-bande*** qui ressemblent très fortement à ceux d'ondelettes de largeur constante en échelle log. Donc, le signal auditif est le résultat d'une convolution avec ces filtres d'ondelettes (et de largeur constante à basse fréquence).

Les algorithmes procèdent selon les étapes suivantes:

- après avoir extrait les échantillons dans un intervalle de largeur N, on calcule l'énergie en fréquence pour chaque bandes critiques (MEL filter bancs[92] comme illustré sur la figure 26.
- On va coder la composante m du signal dans la base DCT ($\check{x}[m]$) telle l'erreur D_m de codage ne soit pas perceptible. Or, le seuil de perception est fonction de l'énergie dans chaque bande. Comme $D_m \propto \Delta_m^2$ donc, le pas de quantification Δ_m est calculé en fonction de l'énergie pour chaque bande critique.

Donc finalement, il semble bien que l'on ajuste l'erreur au problème de perception avec un

91. Voir par exemple http://www.cochlea.eu/son/psychoacoustique. Voir aussi Cours 2020 Sec 7.3 *Digression naturaliste.*

92. Voir Cours 2020 Sec. 7.4 MFC (Mel-Frequency Cepstrum).

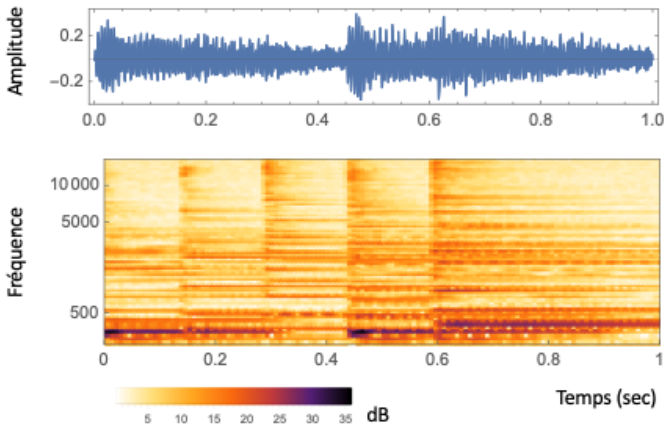

FIGURE 26 – Exemple de calcul d'énergie par bandes de fréquences critiques et par intervalle de temps de 1024 échantillons.

pas variable. Or, nous avons vu la séance dernière que l'utilisation de normes quadratiques pondérées rend compte tout aussi bien du phénomène. C'est alors les poids de pondération qui sont ajustés par l'énergie dans les bandes critiques. Ce sont des poids non fixés à l'avance mais calculés selon le signal en présence.

En pratique pour l'audio $[0, 20k]$Hz, on définit 25 bandes critiques et les standards permettent d'obtenir des compressions qui donnent un débit de 100kbits/sec soit environ un facteur 7 de compression tout en gardant une excellente qualité d'écoute.

9.4 Le cas de l'image: standard JPEG

Il nous faut de nouveau adapter la base orthonormale. Comme dans le cas du son dans la section précédente, on commence par découper l'image $N \times N$ en blocs de $L \times L$ pixels avec typiquement $L = 8$ pour JPEG et $L = 16$ en video. Si on a des imagettes dont l'intensité est régulière, il est naturelle de décomposer dans une base de cosinus dans les 2 directions (u, v) de l'image. On va alors utiliser le petit théorème suivant

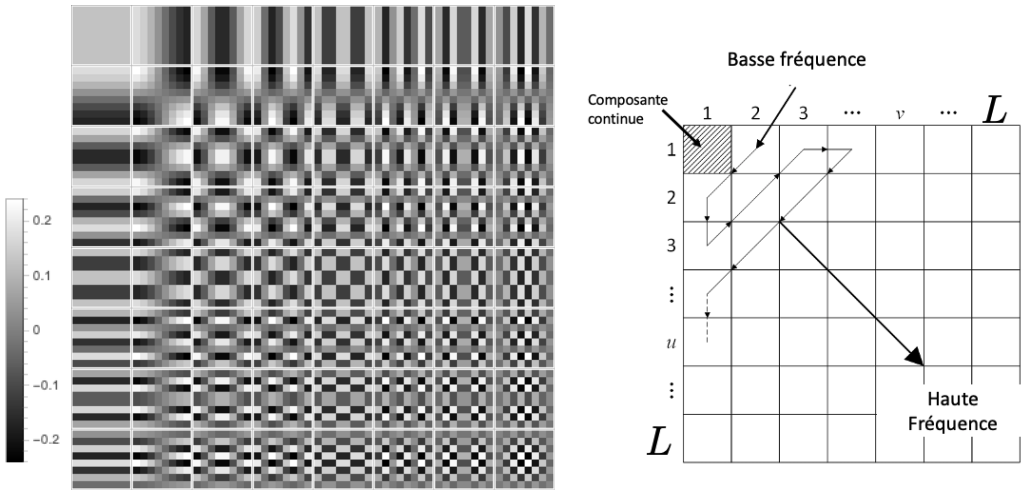

FIGURE 27 – Les 64 éléments de la base de cosinus 2D (Eq. 292) avec $L = 8$. On passe des basses fréquences aux hautes fréquences en se déplaçant de haut à gauche vers le bas à droite selon un *zigzag ordering*. La couleur du carré $m_1 = m_2 = 0$ est donnée par la valeur constante des pixels valant 1/8.

Théorème 23

Si $\mathscr{B} = \{g_m[n]\}_{m<L}$ *est une base orthonormée de* $\mathbb{R}^L$*, on peut obtenir une base orthonormée séparable de* $\mathbb{R}^L \times \mathbb{R}^L$ *en faisant le produit*

$$\{g_{m_1,m_2}[n_1,n_2] := g_{m_1}[n_1]g_{m_2}[n_2]\}_{(m_1,m_2)<L} \tag{291}$$

Ainsi, les éléments de la base 2D se lisent

$$\mathscr{B}_{cos,2D} = \left\{ g_{m_1,m_2}[n_1,n_2] = \lambda_k \lambda_j \frac{2}{L} \cos\left(\frac{\pi m_1}{L}(n_1+1/2)\right) \cos\left(\frac{\pi m_2}{L}(n_2+1/2)\right) \right\}_{(m_1,m_2)<L} \tag{292}$$

avec les λ_k, λ_j définis précédemment pour la base de cosinus 1D. Le couple (n_1, n_2) identifie un pixel dans l'imagette de taille $L \times L$. Les 64 éléments de la base 2D avec $L = 8$ sont donnés sur la figure 27. L'imagette extraite de l'image initiale est décomposée dans cette

base de produit de cosinus selon

$$x[n_1, n_2] = \sum_{m_1,m_2} \langle x, g_{m_1,m_2} \rangle \; g_{m_1,m_2}[n_1, n_2] \tag{293}$$

Notons que l'on peut regrouper le couple d'indices (n_1, n_2) dans un indice unique n et l'on retrouve le type d'expression utilisée en 1D. Et comme en 1D nous pouvons définir une base orthogonale sur l'ensemble de l'image par déplacement d'une fenêtre 2D glissante.

Maintenant, ***l'objectif est d'obtenir une représentation parcimonieuse du signal*** (ici l'image), la question qui se pose alors est de savoir quand est-ce que l'on a des grands coefficients? Un exemple est donné sur la figure 28. On constate clairement que dans les deux cas, le coefficient qui correspond à $m_1 = m_2 = 0$ est le plus grand, c'est celui de plus basse fréquence qui mesure la somme des pixels de chaque image à un facteur prés [93]. Mais à part ce coefficient, dans l'image randomisée, les coefficients sont proches de 0, par contre pour l'image qui présente 2 régions uniformes, on a 2 ou 3 autres coefficients d'amplitudes non nulles. Donc, ***on a plus de coefficients non nuls quand il y a des transitions/discontinuités***. Ceci dit on aimerait réduire autant que faire ce peut le nombre de ces coefficients non nuls. Donc, il faudrait des tailles d'imagettes ajustées pour ne par rencontrer des transitions d'intensité. Pourquoi ne pas réduire à $L = 2$ la taille? La raison est que l'on ne peut obtenir des fréquences plus basses que la taille de l'imagette, donc pour une plage uniforme dont la taille est assez grande, on ne va pas rendre compte de la redondance entre imagettes et l'on va coder le coefficient de basse fréquence plusieurs fois inutilement (Fig. 29). En revanche, si les imagettes sont trop grandes, alors chacune risque de contenir une discontinuité ce qui engendre plusieurs coefficients non nuls à coder. ***Donc, il nous faut prendre les imagettes les plus grande possibles avec des compromis pour obtenir le meilleur taux de compression.*** Après des tests, il semble que pour des images typiques prisent dans la vie courante, la taille de 8×8 soit un bon compromis.

On peut indexer les coefficients pour obtenir une progression des basses fréquences vers les hautes fréquences (*zigzag ordering*), et donner l'équivalent d'un spectrogramme comme sur la figure 30. On constate bien la décroissance très rapide de l'amplitude des coefficients en fonction de la "fréquence" (index du coefficient). Ainsi, ***les informations retenues pour le codage*** sont les suivantes:

93. nb. Avec la définition prise de la base le facteur est $1/L$. Pour une image dont la valeur moyenne d'un pixel est 128, alors la valeur du premier coefficient est $128 * L \approx 1024$.

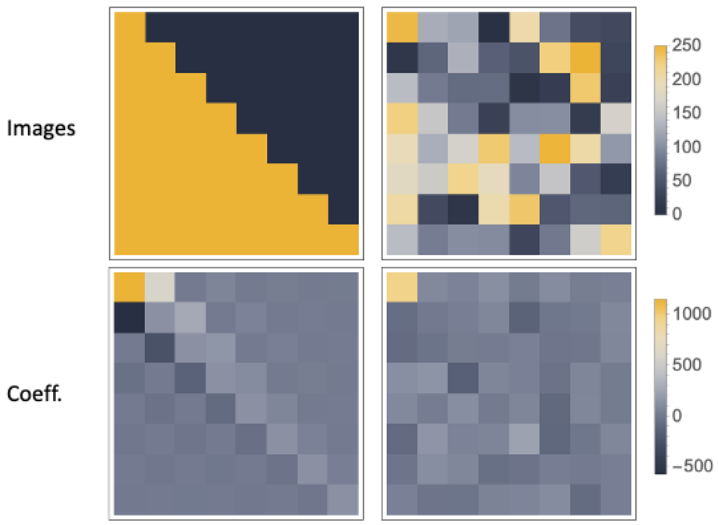

FIGURE 28 – Exemples de calcul des coefficients de décomposition d'imagettes dan la base des 64 éléments de la figure 27. En haut: les images 8×8 dont l'intensité est codée sur 8bits, avec à gauche une image de transition entre 2 régions uniformes et à droite une image random uniforme. En bas, le tableau des coefficients pour chaque image. Les échelles sont communes soit pour les images, soit pour les coefficients.

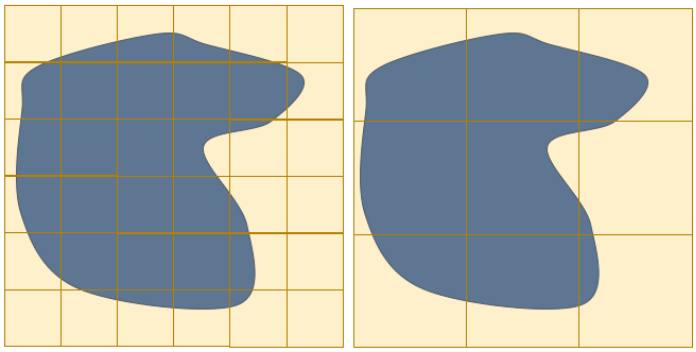

FIGURE 29 – Exemple de 2 pavages par des imagettes de tailles différentes de la même image sous-jacente (en bleu). Si les imagettes sont trop petites, le cas à gauche, alors l'unique coefficient de basse fréquence est répété plusieurs fois (9 environ) inutilement. Si les imagettes sont trop grandes (cas à droite), alors il y a des discontinuités dans chacune d'entre-elles et donc par imagette il y a plusieurs coefficients non nuls à coder, donc une perte du taux de compression.

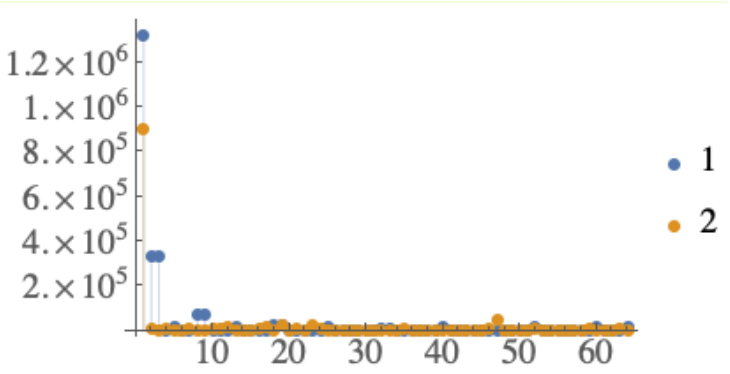

FIGURE 30 – Les valeurs du module au carré des 64 coefficients de la décomposition des 2 images de la figure 28: "1" correspondant au cas de l'image présentant une discontinuité, et "2" pour le cas de l'image random.

FIGURE 31 – A gauche la compression JPEG est de 0.2bpp, à droite de 0.5bpp.

- ***la position des coefficients dont la quantification est non nulles***. Ce qui se fait d'abord par l'intermédiaire d'un vecteur binaire de taille L^2 où chaque 1 indique l'index du coefficient dont la quantification est non nulle. Ensuite, on comprime ce vecteur par un codage dit *Run Length Code* qui code la longueur des portions de même valeur, I pour les 1 et Z pour 0, avec un code entropique, puis on termine par un mot particulier *end-of-block* qui signifie que l'on a que des 0.
- ***la valeur quantifiée de ces coefficients***.

Il y a des librairies qui permettent de compresser au format JPEG en donnant un facteur de qualité comme `ImageMagick` ou `cjpeg` sous Linux[94]. Il faut donc procéder à un petit calibrage afin de faire la correspondance entre ce facteur et le nombre de bits par pixel. Sur la figure 31 sont présentés deux niveaux de compression (0.2bpp et 0.5bpp) de la même image de taille 512×512 codée initialement à raison de 8bits/pixel (bpp). Contrairement à ce que l'on pourrait penser en première intention, ***on peut restituer l'image avec une bonne qualité visuelle même avec un nombre de bits par pixel plus petit que 1***.

Pourquoi est-ce possible? La raison profonde est que l'on a utilisé la redondance spatiale par l'usage de la transformée orthogonale. Si on avait considéré les pixels indépendants les uns des autres on n'aurait pu que binariser l'image ce qui correspondrait

94. https://imagemagick.org/script/convert.php, https://www.unix.com/man-page/linux/1/cjpeg/

à coder chaque pixel soit avec la valeur 0 ou 1 (1bpp)[95]. Le gain de la transformée est d'au moins un facteur 2. Notez que si l'on zoom sur l'image, on observe des phénomènes d'oscillations (*phénomène de Gibbs*[96]) car on a enlevé les hautes fréquences.

Maintenant, changer ***le taux de compression se fait en changeant le pas de quantification***. Attention, la base des produits de cosinus est la même, c'est-à-dire que L est toujours de 8 pixels. En audio, on a été plus subtile en utilisant une norme pondérée pour s'adapter à la perception de la l'oreille. Donc, si l'on passe d'un pas égal Δ_1 à un pas plus grand $\Delta_2 > \Delta_1$, tous les coefficients plus petits que Δ_2 qui n'étaient pas nuls avec Δ_1 sont mis à zéro, on dégrade donc la qualité de restitution de l'image en détruisant les hautes fréquences, ce qui donne des ***effets de blocs***. C'est un peu mieux au niveau des mats du bateau, car l'intensité est plus grande donc il y a plus de bits pour représenter le signal dans ces cas. L'effet est donc beaucoup plus visible dans les régions de l'image où il n'y a pas beaucoup de structures comme le ciel.

Comment obtenir des taux de compression plus importants sans détruire la qualité de l'image? Il nous faut utiliser différents niveaux d'échelle et exploiter les redondances à tous les niveaux. Cela se fait ***à l'aide de bases d'ondelettes orthonormales***.

9.5 Usage des Ondelettes: standard JPEG2000

NDJE: pour l'introduction des bases d'ondelettes 1D et 2D, il n'aurait pas été opportun de faire un copier-coller du cours de 2021. Je vous revoie donc par exemple aux sections 5.3, 6.3, 8 puis 9.3. Vous pourrez également trouver d'autres informations dans le Cours de 2020.

Juste pour mémoire, la décomposition en ondelettes se fait par un algorithme rapide en bancs de filtres de complexité $O(N)$ plus rapide que la FFT. ***Les seuls coefficients non nuls sont ceux pour lesquels l'ondelette localisée à la fois en fréquence et dans l'espace tout en respectant l'inégalité de Heisenberg, signale la présence d'une discontinuité*** comme on peut le constater sur la figure 32. Ensuite, on procède à une quantification des coefficients non nuls un peu à la façon de JPEG. Le résultat de la figure 33 montre

95. Voir Cours 2021 Figure 57.
96. Voir Cours 2021, note de bas de page Sec. 8.3

FIGURE 32 – Exemple de décomposition en ondelettes d'une image de bateau. On a renversé la colormap pour faire apparaitre les coefficients non nuls qui sont peu nombreux et localisés aux endroits des discontinuités de teinte.

l'efficacité de la décomposition en ondelettes pour un taux de compression de 0.2bpp en comparant des images JPEG et JPEG2000 [97].

Si l'on pousse d'un facteur 4 supplémentaire la compression (Fig. 34) pour obtenir un taux de 0.05bpp, on commence à voir des distorsions. Ce ne sont pas des effets de blocs comme pour JPEG, mais plutôt le résultat du manque de coefficients à hautes fréquences qui s'apparente au phénomène de Gibbs présent en JPEG. Cependant, les effets sont moindre en JPEG2000.

97. NDJE: Les images ont été obtenues à partir en utilisant l'outil convert/ImageMagick sous Linux/Mac en ajustant l'option "-define jp2:rate=x" en plus des options "jp2:nomct" et "jp2:numrlvls=4". Ensuite, *pdflatex* ne comprenant pas le format "jp2" j'ai utilisé une conversion en "png" des fichiers, et vérifier que le rendu était le même.

FIGURE 33 – Exemples de différence de qualité d'images restituées avec le même taux de compression de 0.2bpp soit en utilisant le standard JPEG à gauche, ou JPEG2000 à droite. On voit que pour des taux de compression assez grand, l'utilisation des ondelettes permet de tenir compte des redondances à toutes les échelles afin d'obtenir une meilleure restitution.

FIGURE 34 – Images au format JPEG2000 avec un taux de 0.05bpp.

9.6 Confrontation de la théorie à un cas réel

Après avoir montré des exemples de compression JPEG et JPEG2000 voyons ce qu'il en est de l'adéquation avec la théorie et en particulier l'expression de la distorsion D en fonction du nombre de bit par pixel $\bar{R}$ du théorème 22. Pour cela, nous introduisons un indicateur de qualité, le PSNR (*peak signal-to-noise ratio*). Pour des images comportant N pixels dont la valeur est codée sur 8bits (valeur maximale d'un pixel est égale à 255)

$$\boxed{PSNR(\bar{R}, \overline{\mathbb{H}}_d) := 10 \log_{10} \frac{255^2}{D(\bar{R}, \overline{\mathbb{H}}_d)/N}} \tag{294}$$

où $\overline{\mathbb{H}}_d$ est donné pour refléter la dépendance vis-à-vis des propriétés du signal. Donc, l'on s'attend à ***une relation linéaire*** du type

$$PSNR(\bar{R}, \overline{\mathbb{H}}_d) = (20 \log_{10} 2)\bar{R} + C(\overline{\mathbb{H}}_d) \tag{295}$$

Sur la figure 35 on montre l'évolution du PSNR en fonction de $\bar{R}$ pour la figure du bateau que l'on a comprimé avec les formats JPEG et JPEG2000. Ce que l'on constate est bien

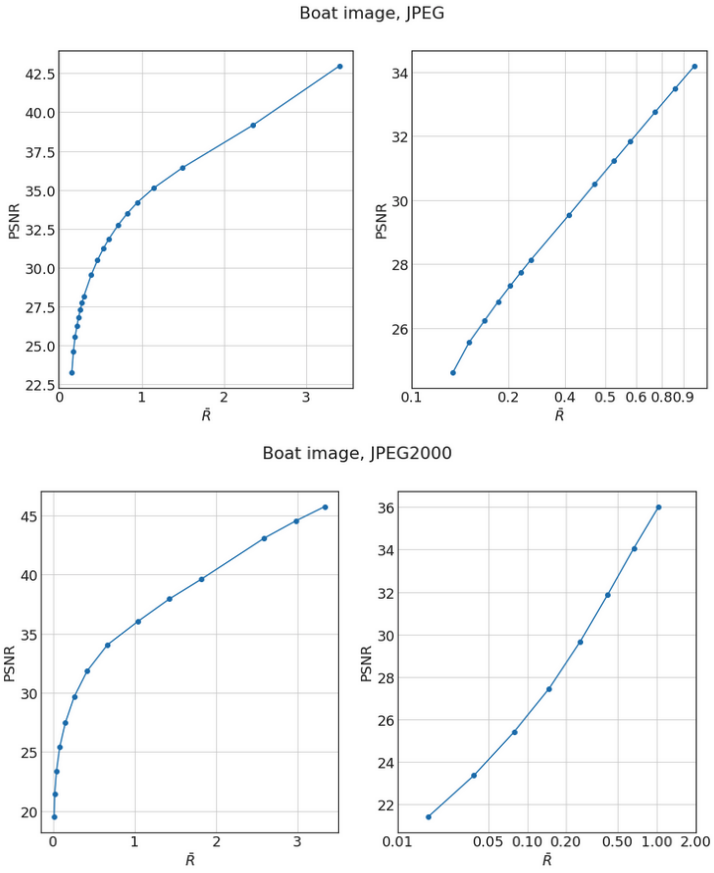

FIGURE 35 – Évolution du PSNR en fonction du nombre de bits par pixels pour l'image du bateau compressée au format JPEG (en haut) et JPEG2000 (en bas) . A droite, l'échelle de $\bar{R}$ est en $\log_2$, alors qu'à gauche elle est linéaire.

un comportement linéaire[98] pour $\bar{R} > 1$ par contre comme le montre la figure de droite, ***pour*** $\bar{R} < 1$ ***le comportement est plutôt linéaire en*** $\log_2(\bar{R})$. Il faut donc amender la théorie et pour cela comprendre d'où vient ce phénomène de perte de qualité d'image.

Pourquoi est-il important de s'intéresser à la zone $\bar{R} < 1$? En pratique, c'est la région d'intérêt comme on l'a expliqué à la suite de la démonstration du théorème de Shannon (Th. 16). La raison vient de la distribution des coefficients d'ondelettes par

98. NDJE: à l'heure où sont écrites ces notes ces courbes sont encore préliminaires, car que cela soit avec `ImageMagick/convert` ou bien `cjpeg`, la pente n'est pas celle escomptée de $20\log_{10} 2$ mais environ un facteur 2 en moins pour $\bar{R} > 1$.

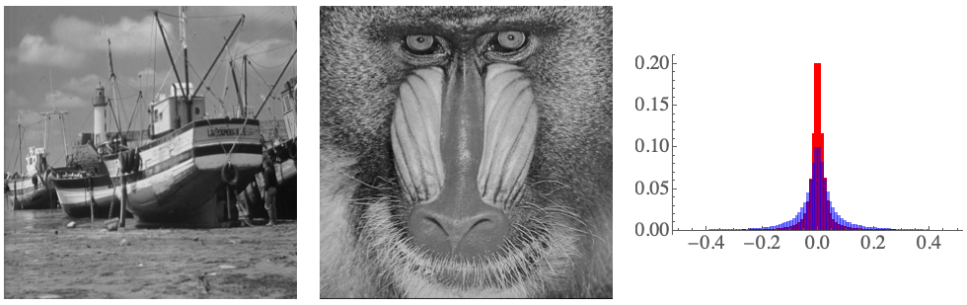

FIGURE 36 – A gauche et au milieu deux images. A droite, en rouge l'histogramme normalisé des coefficients d'ondelettes (de détails seulement) pour l'image du bateau, en bleu pour celle du singe montrant plus de textures.

exemple donnée sur la figure 36. Ce que l'on remarque c'est que ces distributions sont très piquées en 0, ce qui met à mal ***l'hypothèse de haute résolution*** qui demande une distribution constante sur une échelle Δ. L'erreur est donc la quantification dans le bin $[-\Delta/2, \Delta/2]$.

9.7 Comportement quand $\bar{R} < 1$

Nous allons faire le lien avec le cours de 2021. Prenons un signal auquel on applique une quantification sur les coefficients de décomposition dans une base orthonormale:

$$\hat{x} = \sum_m Q(\langle x, g_m \rangle)\ g_m \tag{296}$$

La distorsion est donnée par

$$\begin{aligned} D &= \sum_m |\langle x, g_m \rangle - Q(\langle x, g_m \rangle)|^2 \\ &= \sum_{|\langle x, g_m \rangle| \leq \frac{\Delta}{2}} |\langle x, g_m \rangle|^2 + \sum_{|\langle x, g_m \rangle| > \frac{\Delta}{2}} |\langle x, g_m \rangle - Q(\langle x, g_m \rangle)|^2 \end{aligned} \tag{297}$$

Soit M le nombre de coefficients dont l'amplitude est plus grande que $\Delta/2$. En reprenant un argument de 2021, l'approximation qui consiste à ne prendre que M coefficients de la

décomposition [99]

$$\tilde{x}_M = \sum_{m \in I(M)} \langle x, g_m \rangle \; g_m \tag{298}$$

génère une erreur

$$\|x - \tilde{x}\|^2 = \sum_{m \notin I(M)} |\langle x, g_m \rangle|^2 \tag{299}$$

Si on veut rendre cette erreur la plus petite possible, il faut que les produits scalaires en valeur absolue soient les plus petits pour $m \notin I(M)$, et donc soient les plus grands pour $m \in I(M)$. En particulier, on fixe un seuil $T(M)$ qui ne garde que les M plus grands coefficients pour définir $I(M)$:

$$I(M) = \{m \; / \; |\langle x, g_m \rangle| > T(M)\} \quad \text{et} \quad T(M) \; tq. \; |I(M)| = M \tag{300}$$

Donc, en fixant le seuil à $T(M) = \Delta/2$, on fixe non seulement le nombre M de coefficients retenus mais aussi le niveau d'erreur d'approximation. Appelons x_M l'approximation de x telle que

$$x_M = \sum_{|\langle x, g_m \rangle| \geq \frac{\Delta}{2}} Q(\langle x, g_m \rangle) \; g_m \tag{301}$$

(la somme comporte que M termes), il vient alors que la distorsion D peut être encadrée selon

$$\boxed{\|x - x_M\|^2 \leq D \leq \|x - x_M\|^2 + M\frac{\Delta^2}{4}} \tag{302}$$

En effet, l'erreur $\|x - x_M\|^2$ est donnée par la puissance des coefficients pour lesquels $|\langle x, g_m \rangle| \leq \Delta/2$ et une majoration de $|\langle x, g_m \rangle - Q(\langle x, g_m \rangle)|$ est $\Delta/2$.

Pour obtenir une relation de type $D(\bar{R})$, il nous faut relier le nombre de coefficients M à $\bar{R}$, et $\|x - x_M\|$ à M. Afin de parvenir à faire ce schéma, il nous faut une hypothèse sur les produits scalaires, laquelle est guidée par les profils de leurs distributions en pratique.

99. NDJE: $I(M)$ est un ensemble qui en Fourier serait par exemple les M coefficients de basse fréquence (approche linéaire: M ne dépend pas du signal), en Ondelettes cela serait l'ensemble des coefficients dont l'amplitude est plus grande qu'un certain seuil (effet non linéaire: car le seuil dépend du signal lui-même).

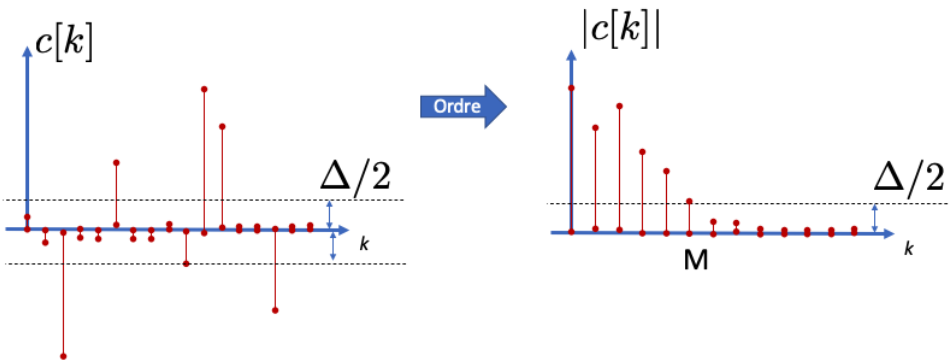

FIGURE 37 – Illustration de la mise en ordre des coefficients $c[k]$ et relation entre le nombre de coefficients M et le seuil $\Delta/2$.

Théorème 24

Si l'on ordonne les coefficients de la décomposition de x dans la base par ordre décroissant (Fig. 37)

$$c[k] = |\langle x, g_k \rangle|, \quad c[k] \geq c[k+1] \tag{303}$$

et que l'on suppose que la décroissance est de la forme

$$|c[k]| = ck^{-\alpha} \quad \alpha > 1/2, \ c > 0 \tag{304}$$

alors la distorsion a le comportement suivant

$$\boxed{D(R) \approx \left(\frac{R}{(\alpha - 1)\log_2 R + O(\log N)} \right)^{1-2\alpha}} \tag{305}$$

Démonstration 24. Nous allons donner l'architecture de la démonstration. Dans un premier temps envisageons de calculer $\|x - x_M\|^2$, il vient d'après l'ordre des coefficients et le fait que x_M garde que les M premiers coefficients (les plus grands)

$$\|x - x_M\|^2 = \sum_{k=M+1}^{N} |c[k]|^2 \approx \sum_{k=M}^{N} c^2 k^{-2\alpha} \approx c^2 M^{1-2\alpha} \tag{306}$$

Maintenant, il nous faut relier M à R le nombre de bits. Pour ce faire on distingue dans R, la contribution R_0 qui code la position des coefficients non nuls, et la contribution R_1 qui code leurs valeurs. Or, les coefficients non-nuls satisfont la relation de décroissance, donc a priori $c[k] \in]-c, c[$, intervalle qui est découpé en boites de largeur Δ. Donc, le nombre de boites de quantification est

$$K \approx \frac{2c}{\Delta} \tag{307}$$

et $\log_2 K$ donne le nombre de bits d'un codage grossier d'un coefficient. Ainsi

$$R_1 = M \log_2(2c/\Delta) \tag{308}$$

Maintenant, la relation entre M et Δ est donnée à peu de chose près par

$$c[M] = \frac{\Delta}{2} \approx cM^{-\alpha} \Rightarrow R_1 \approx \alpha M \log_2 M \tag{309}$$

Pour obtenir R_0, il nous faut calculer l'entropie d'une variable aléatoire qui prend la valeur 1 avec une probabilité p et la valeur 0 avec la probabilité $1-p$. Il s'agit de l'entropie d'une loi de Bernoulli

$$\mathbb{H}_B = -(1-p)\log_2(1-p) - p\log_2(p) \tag{310}$$

Or, $p = M/N$ la fréquence d'apparition des coefficients non-nuls parmi les N coefficients de la décomposition. Ainsi [100], $R_0 = N\mathbb{H}_B$. Finalement, dans le cas d'un fort taux de compression $M/N \ll 1$, alors

$$R/M \approx (\alpha - 1)\log_2 M + \log_2 N + o(\log_2 N) \tag{311}$$

L'inversion de la relation donne M en fonction de R qui donne

$$M \approx \frac{R}{(\alpha-1)\log_2 R + \log_2 N + o(\log_2 N)} \tag{312}$$

100. nb. il y a autant de coefficients que d'échantillons du signal, soit N.

Et comme $M(\Delta/2)^2 \approx c^2 M^{1-2\alpha}$, tout comme $\|x - x_M\|^2$, alors $D(R)$ suit bien une loi

$$D(R) \approx \left(\frac{R}{(\alpha - 1)\log_2 R + \log_2 N + o(\log_2 N)} \right)^{1-2\alpha} \tag{313}$$

■

Ce qui va être très important à fort taux de compression ($\bar{R}$ petit), c'est bien le nombre de coefficients non nuls M, car le nombre de bits lui est *grosso modo* proportionnel tout comme l'erreur, laquelle doit être la plus petite possible. ***Et finalement, une fois que l'on sait faire du codage efficace, le plus difficile c'est de trouver la base qui va comprimer au mieux l'information.*** Cette base optimale élimine la redondance des données en trouvant des formes de régularité qui était l'objet du cours de 2021 en parcourant le triangle: ***Régularité, Approximation, Parcimonie***.

10. Conclusion

Finalement, un siècle après la parution de l'article de Fisher, nous sommes toujours dans le cadre de son programme d'essayer de trouver un modèle paramétrique de la distribution qui reflète au mieux les observations. Ceci est au cœurs du Machine Learning, et d'une certaine manière les réseaux de neurones sont des systèmes paramétrés conçus pour *in fine* maximiser la vraisemblance, le tout en effectuant une descente de gradients pour déterminer les paramètres. Bien entendu, ce qui a fondamentalement changé en un siècle, c'est la complexité des modèles.

Cet aspect probabiliste est selon S. Mallat fondamental pour comprendre les résultats si déroutant des réseaux de neurones. Disons que le point de vue probabiliste, par rapport au point de vue déterministe, est bien le cadre conceptuel dans lequel on un espoir de comprendre les propriétés statistiques, car finalement si ces réseaux arrivent à estimer ces distributions paramétrées en très grande dimension, c'est qu'ils arrivent à capturer une information très pertinente contenue dans les observations, information qui semble simplifier grandement le problème. Or, cette simplification est sans doute liée à la concentration de la probabilité qui est une transcription du fait que les observations ne sont pas quelconques, elles "vivent" dans des ensembles typiques de Shannon.

Entre les années 1950 à 2000 environ, le cadre de la Théorie de l'Information a exploité pleinement en particulier l'usage de l'entropie. Mais le problème cependant qui s'est fait jour, c'est qu'à part quelques cas simples à base de gaussiennes pour faire court, on ne savait pas calculer l'entropie des systèmes. Et c'est là qu'à partir des années 2000, et de plus en plus avec des réseaux de neurones à nombre colossal de paramètres, on arrive à beaucoup mieux appréhender ces ensembles typiques des distributions de probabilités. Donc, si le cadre théorique reste le même, la (bonne) surprise c'est que l'on s'approche de ces ensembles typiques, ce qui permet d'attaquer de nouveaux problèmes comme rentrer dans le détails des erreurs d'approximation par exemple à l'aide de l'analyse harmonique.

Maintenant, le point de vue de Shannon qui ignore toute forme de paramétrisation pour ne s'intéresser qu'à l'information intrinsèque des observations, et le point de vue de Fisher, sont finalement en relation naturelle. En Physique Statistique particulièrement qui est confrontée à des observables dont on tire des moments (moyenne au sens large), si l'on veut inférer des distributions en maximisant l'entropie (Principe d'Entropie Maximale de Jaynes), cela revient à des modélisations par la famille exponentielle, donc à un formalisme de probabilités paramétrées à la Fisher. Parmi les questions qui se posent par exemple on peut ce demander si on peut faire mieux dans le cadre des compressions de données avec les réseaux de neurones. Nous verrons peut être dans un avenir proche de nouveaux standards apparaître.