Notes et commentaires au sujet des conférences de S. Mallat du Collège de France (2020)

Jean-Eric Campagne

Notes et commentaires au sujet des conférences de S. Mallat du Collège de France (2020)

Modèles multi-échelles et réseaux de neurones convolutifs

J.E Campagne *

Fev 2020; rév. 3 novembre 2023

*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.

Cette année 2020 c'est la troisième du cycle de la chaire de la Science des Données de S. Mallat, le thème en est: **Modèles multi-échelles et réseaux de neurones convolutifs**.

Notons enfin que cette année, les cours ont été écourtés à la suite de la décision de procéder à un confinement général pour contrer la progression de la pandémie Covid-19 SARS-2. En présentiel, le cours de S. Mallat s'est achevé le 11 Mars 20. Le dernier cours a été enregistré le 15 Juin 2020 après la période de confinement.

2. Séance du (22 Janv.)

2.1 Introduction thématique du cours

Une série de questions se posent autour des systèmes d'apprentissage d'un point de vue mathématiques appliquées et en particulier cette année au sujet les réseaux de neurones.

2.1.1 Comprendre un système d'apprentissage?

Qu'entend-on par *comprendre* dans le cadre de Math. Appli.? En effet, selon les ingénieurs qui les utilisent, ils comprennent complètement les algorithmes qu'ils mettent en oeuvre ainsi que les architectures des réseaux qui sont complètement spécifiées. D'un autre côté, on ne comprend pas vraiment le **"pourquoi ça marche?"**. C'est-à-dire que l'on ne comprend pas par exemple : ni les performances de généralisation, ni quand est-ce que le réseau va donner une très bonne précision pour n'importe quel example pour lequel on estime une fonction $f(x)$. Donc, la notion de compréhension au sens mathématique n'est pas la même que celle au sens ingénierie par exemple.

Ceci étant dit, quand on commence à étudier des sujets tels que l'audition et la vision d'un point vue physiologique, on se rend vite compte de la merveille de la Nature qui a développé des systèmes très sophistiqués partant des récepteurs vers le centre névralgique du traitement d'information et vice-versa. Au bout d'un moment face à la complexité du système, on se dit que c'est bien trop compliqué et qu'il faut se concentrer sur l'apprentissage lui-même qui semble plus abordable. Bien entendu cela paraît très frustrant et par certains cotés insatisfaisant de ne pas rentrer dans les détails du fonctionnement de ces systèmes. Mais, l'espoir est que cela soit plus simple, or on va le voir il n'en est rien. Ce n'est pas plus simple du tout de comprendre cette boucle d'apprentissage surtout quand on veut creuser le problème au-delà de l'algorithmique.

D'un point de vue très pratique, il y a des problèmes :

— de **robustesse**: souvent les systèmes ne sont pas très stables et induisent des erreurs tout à fait dramatiques (ex. les applications dans la voiture autonome, dans le médical, etc). Comment améliorer cette robustesse?

— d'**efficacité**: le nombre de données qu'il faut pour l'apprentissage, l'énergie qu'il faut dépenser pour faire cet apprentissage en un temps raisonnable (cf. ferme de GPUs), sont tout à fait prohibitifs. Est-ce optimal ou pas du tout?

— de **contrôle**: l'architecture va permettre d'apprendre un certain type de problème, mais le temps pour faire le lien entre les spécifications de l'architecture et le type de problème représente 99% du temps de design.

Ce dernier aspect va être au centre du cours et on va le voir à partir de la notion **d'information *a priori***. De quelle information *a priori* dispose-t'on, comment peut-on exprimer cette information d'un point de vue mathématique. Donc, ce que l'on va voir au cours des séances futures c'est :

1. quel est le lien entre **l'architecture** et la **généralisation**,
2. que ces outils (ex. NN [1]) peuvent être utilisés comme **outils d'exploration de la complexité**. C'est un thème qui dépassera ce cours. Par exemple, on prend un système d'apprentissage (NN, DT, méthode à noyaux...), on essaye de calculer des fonctions physiques (ex. énergie d'un système) et on voit si ça marche. Si tel est le cas, cela nous apprend quelque chose de la complexité de la fonctionnelle sous-jacente. Ex. des équipes en chimie quantique calculent l'énergie de molécules ce qui permet d'accéder à leurs stabilités ou autres propriétés, et donc d'accéder aux

1. On notera: NN pour Neural Network, DT pour Decision Tree, RF pour Random Forest

propriétés des matériaux. Si donc on est capable de faire cela avec un NN, ensuite on peut revenir sur la physique et se poser des questions sur le pourquoi de telle ou telle propriété de la molécule. Actuellement, on est complètement dans l'empirique, il y a des systèmes qui marchent, mais on ne comprend pas le lien avec l'équation de Schrödinger. On assiste donc à un nouveau regard sur la chimie, un peu similaire à la physique statistique, où l'on regarde le système dans son ensemble. On identifie des sortes de "macro-variables" lesquelles sont calculées par ces réseaux et qui sont capables de spécifier l'énergie de la molécule. La question est de savoir "Pourquoi?". La complexité est sans doute moins grande que ce à quoi on pensait. Bref, c'est en ce sens que les système d'apprentissage sont des outils d'exploration du complexe.

3. de comprendre l'apprentissage de ces architectures de réseaux. Ici, à force de passer beaucoup de temps au design des architectures de réseaux, on en vient à se poser la question de savoir si finalement on ne pourrait pas "**apprendre ces architectures**". Il y a un certain nombres de méta-paramètres qui définissent ces architectures (ex. nombre de couches, nombre de neurones par couche, type de couche...) et donc pourquoi ne pas envisager un système d'apprentissage de haut niveau qui apprendrait à trouver la meilleure architecture pour telle ou telle application. On pourrait penser à un algorithme génétique (darwinien) qui va sélectionner la bonne solution à force de transformations/mutations du système.

Le dernier point est un niveau de questionnement que l'on n'abordera pas cette année mais il fait parti du paysage de l'ensemble des questions que l'on peut aborder.

2.1.2 Le rôle des mathématiques

Avant d'aller plus loin, on peut aborder quel est le rôle des mathématiques dans cette problématique de l'apprentissage. En effet, on a un regard sur la linguistique (voir le cours de 2019), et maintenant on a des systèmes qui sont capables de faire de la traduction, de l'analyse de texte pour reconnaitre l'auteur(e), de faire de la génération de texte, etc. On a une sorte de mise en abîme, où on se pose la question pourquoi le langage mathématique est aussi efficace à rendre compte de phénomène de la Nature. Le point de vue que l'on prend ici, est celui de maths appliquées, et ce langage évolue au fur et à mesure des questions que l'on se pose. En mathématiques "pures" la problématique se situe au sein du champ des mathématiques, tandis qu'en mathématiques "appliquées" les questions viennent de

l'extérieur du champ disciplinaire. Ici en quelque sorte, le langage des math. appli. évolue comme un système d'apprentissage où les données (input) sont les problèmes posés. En ce sens d'ailleurs jusqu'à la fin du XIXe siècle en gros, les questions mathématiques étaient sous-tendues des problèmes de physique et il est donc pas étonnant que les mathématiques soient remarquablement adaptées à décrire des phénomènes physiques.

Dans le cas qui nous intéresse, les problèmes sont très complexes et les mathématiques ont donc besoin d'évoluer. Le champ disciplinaire en action ici est celui de la **très grande dimension**, c'est-à-dire qu'il manipule des fonctions à très grand nombre de variables. il est très transversale au sein des mathématiques d'aujourd'hui: probabilité, analyse (harmonique) [2], etc... Les questions dans ce domaine de la très grande dimension sont ouvertes et donc la recherche est très active pour faire évoluer les mathématiques.

2.1.3 Principes d'organisation des architectures

Le problème, comme on l'a vu par exemple dans le cours de 2018, c'est la grande dimension: une image typique a des millions de pixels, de même un son a des millions d'échantillons par seconde, un texte a un million de caractères, et pire encore une molle a par définition 10^{23} entités, par cela la Physique/Chimie a toujours été la science de la très grande dimension... Comment aborder ce problème? En fait, on va mettre en œuvre 4 principes: **la séparabilité, les symétries, et la parcimonie** [3] et un quatrième que l'on peut qualifier de meta-principe qui est **l'évolution**.

Une idée qui vient immédiatement est d'opérer une **réduction de dimension** en utilisant des *a priori*, par exemple en essayant de trouver des formes de **séparabilités**. Ainsi, dans le cas (simple) d'une image, les **interactions entre pixels** sont essentiellement **locales** [4]. Dans ce cas, on peut se contenter de traiter des patchs séparément (attention néanmoins aux jointures) puis de rassembler les résultats individuels pour obtenir une réponse sur le problème initial.

2. En analyse on va se rendre compte que la connaissance a priori est déduite de pas grand chose et que les données sont structurées sur des espaces à faible dimension topologique (le temps, des images), ce qui permet de dérouler tout une analyse.

3. voir par exemple le cours de 2019 Secs. 3.6, 3.7, 3.8.

4. NDJE: on peut concevoir ce raisonnement après avoir enlever une contribution globale qui influencerait l'ensemble d'un champ de vue.

Ceci dit, il y a des problèmes qui nécessitent des structures plus importantes pour lesquelles on ne dispose pas de séparabilité locale. Mais dans la plus part des cas on peut mettre en jeux une **séparabilité entre les diverses échelles** du problème (cf. **hiérarchie**). Cette séparabilité d'échelle est très classique en Physique [5] et permet une réduction de la dimension du problème, si bien que même si les interactions entre échelles peuvent le complexifier, il reste résoluble. Notamment, nous verrons les outils de l'analyse harmonique (temps-fréquence/ondelettes).

Autre point, la recherche de **symétries** et un ingrédient fondamental, notamment en Physique des Particules, car elle permet de trouver des **invariants** (cf. théorème d'Emmy Noether en 1915). Par exemple: des symétries spatiales qui entrainent des invariances par translation/rotation/flip. Une fois identifiée la symétrie, on peut réduire encore une fois la dimensionalité du problème. Un exemple simple: un problème 2D invariant par symétrie de rotation est décrit par une fonction pour laquelle l'angle n'apparait plus dans sa liste de variables. Bien entendu, cette réduction va être d'autant plus effective que l'on a suffisamment de symétries. D'un point de vue mathématiques, on va faire appel à la **théorie des groupes**, dont les groupes de Lie pour les symétries continues.

Le dernier principe d'organisation est la **parcimonie**. Notons que dans le domaine de la reconnaissance, celui-ci était dominé par le thème de la reconnaissance de forme (pattern recognition) avant l'ère de l'apprentissage. Ainsi, le principal sujet était la notion de structure [6] qu'il fallait reconnaitre: ex. les oreilles d'un chien/chat, les yeux dans un visage... Ceci dit cette approche se comprend très bien et les applications mises au point par les ingénieurs sont souvent retrouvées après bien des détours par les mathématiciens voulant faire une sorte de *tabula rasa*. Et encore une fois ici, comprendre ces structures permet également de réduire la dimensionalité du problème. Il est caractérisé par un nombre de structures élémentaires qui est en nombre bien plus petit que la dimension initiale, et donc le but est de décomposer le problème sur ces **structures élémentaires**. Un point au passage, cette décomposition en structures élémentaires peut nécessité d'avoir fait une réduction de type "recherche de symétries" au préalable. Il faut donc avoir à l'esprit les différents types de principes d'organisation. Ceci dit, c'est dans le cadre de la découverte des structures que l'apprentissage va jouer son rôle.

5. C'est par exemple le réductionnisme cartésien de 1648 dans son *Traité de l'homme*.

6. Voir en 2019 la Sec. 2.3 qui traite de l'influence de la sémantique de Noam Chomsky notamment dans ce domaine.

Au dessus de tous ces principes, il y a la notion d'**évolution**. En effet, pour l'apprentissage, on se fixe l'objectif de minimiser un risque, ce qui va faire évoluer les paramètres du réseaux vers une solution. Le chemin de la minimisation peut être considéré comme une **variable de temps**. Cette variable est vue de différentes façons selon le cas de figure. Ainsi, pour des problèmes de traitement du signal $x(t)$, le temps est une **indexation du signal** numérisé. Dans ce cas, le temps n'a pas de propriété particulière par rapport à une variable d'espace par exemple.

Dans beaucoup de sujets de Physique, on peut vouloir décrire non pas l'état du système qui est trop complexe, mais plutôt son évolution. Par exemple en Mécanique (Quantique ou pas), l'équation différentielle régie par un Hamiltonien joue un rôle essentielle:

$$\frac{\partial x(t)}{\partial t} = H(t)x(t)$$

Le temps est vu à travers les **opérateurs d'évolution**.

Il y a une troisième façon de voir le temps, à savoir une forme d'indexation d'événements, c'est-à-dire le temps vu sous l'angle de la **coïncidence**. Cela se traduit par exemple quand on veut suivre un objet de couleur particulière lors d'une série de prises de vue, ce qui change fondamentalement c'est l'intensité lumineuse à la frontière de l'objet en question (cf. le fond est inchangé au cours de la prise de vues). Donc, inversement si dans une prise de vues on observe des changements "en coïncidence" de luminosité pour plusieurs pixels de l'image, alors on peut légitimement se poser la question de savoir si l'ensemble de ces pixels appartiennent à une même structure. Donc, l'idée ici est que derrière les coïncidences "temporelles" on va pouvoir détecter des structures. Ce n'est pas anodin, en effet le bébé âgé de quelques jours pour lequel la vision n'est pas encore totalement opérationnelle, est capable de reconnaître son entourage. Selon toute vraisemblance, c'est par le biais de sa main qu'il fait son apprentissage, car par la préhension et son mouvement complexe, la main crée des discontinuités. Dans le cas d'apprentissage des machines, cette notion de coïncidence est un axe de recherche afin de réduire le nombre d'exemples de training. En effet, actuellement le nombre d'exemples atteint typiquement des 100k voir 1M, soit beaucoup trop en comparaison des systèmes naturelles. Donc, il y a du grain à moudre, surtout que pour le moment l'apprentissage automatique est plutôt statique, c'est-à-dire les exemples sont traités à la queue-le-leu assez indépendamment les uns des

autres [7], alors qu'il faudrait introduire de la dynamique et donner du liant.

2.1.4 Information *a priori*

Cela sera un des thèmes principaux de cette année, car en définitive, on ne peut apprendre sans *a priori*. Tout d'abord, l'information *a priori* se traduit en termes de **classes d'hypothèses**. A partir de données, notées x, on veut trouver une fonction f telle qu'elle puisse donner y, à savoir $y = f(x)$. Le problème est donc d'approximer f que l'on ne connait pas. Ainsi, l'hypothèse se traduit typiquement par se définir un ensemble $\mathcal{H}$ pour lequel les éléments $\tilde{f}$ satisfont diverses propriétés, telle que le risque $\mathcal{R}(f, \tilde{f})$ soit petit. Dans des cas simples, le risque est une norme et $\|f - \tilde{f}\|$ est l'erreur d'approximation.

Donc, l'information *a priori* est concentrée sur l'ensemble $\mathcal{H}$ dans lequel on essaye de trouver la solution. Dans ce cadre, les données vont nous aider à trouver la bonne fonction $\tilde{f}$ (nb. à travers la minimisation du risque empirique). Plus l'information *a priori* est forte, plus l'ensemble des solutions est restreint, et plus le nombre de données pour trouver la solution est petit. A contrario, si on ne dispose pas d'information *a priori*, alors l'ensemble $\mathcal{H}$ est tellement grand que l'on se retrouve devant la malédiction de la dimensionalité (cf. Cours de 2018) car on ne pourra jamais avoir suffisamment de données pour trouver la solution.

Comme on l'a vu les années antérieures, expliciter la définition de $\mathcal{H}$ est essentiellement se définir un niveau de **régularité** des fonctions considérés. Cependant, l'information *a priori* reste des modèles, dont il faut s'interroger à un certain moment sur leurs niveaux d'exactitude. Il ne faut ni perdre de vue cet aspect "d'arbitraire", ni oublier que les résultats obtenus sont dans le cadre de ces hypothèses. Également il faut avoir à l'esprit qu'il y a un échange permanent entre la connaissance *a priori* et les résultats (erreur) que l'on obtient qui peuvent induire un retour aux hypothèses pour les restreindre. Tout dépend du nombre d'échantillons/données que l'on a.

Jusque dans les années 2005-10, on utilisait que de l'information *a priori* pour définir les structures du système, et à la fin, on faisait une simple classification linéaire (ex. pour séparer chien et chat). Ce point de vue, nous allons le prendre pour cheminer et voir

7. Mentionnons néanmoins les systèmes à mémoire de type Recurent Neural Networks, LSTM, GRU...

jusqu'où on peut aller, et à quel moment cela ne marche plus (cf. les hypothèses initiales sont trop restrictives ou bien les ensembles $\mathcal{H}$ sont trop grands).

Finalement, le thème du cours sera: **Nature mathématique de l'information *a priori* en grande dimension**. Il sera donc question de ce que les outils standards des mathématiques peuvent nous apprendre à ce sujet. Par ailleurs, il est toujours intéressant quand on aborde un sujet, de se poser la question suivante: 1) de quelle information *a priori* je dispose, 2) j'essaye de me faire un classificateur linéaire qui inclut cette information, 3) je prends un réseau de neurones et je compare les résultats, 4) si le réseaux me donne de meilleurs résultats comment cela se traduit-il en termes d'information que je n'aurai pas su voir *a priori*.

2.2 L'apprentissage supervisé face à la grande dimension

Dans le cours, on ne va s'intéresser qu'aux cas d'apprentissages supervisés. On se place donc dans un espace des données de très grandes dimension, $x \in \mathbb{R}^d$ avec d très grand (cf. $d \approx 10^{6-9}$). On va étudier deux types de problèmes classiques :

- celui de la **classification**, où la classe des labels est donnée par $f(x)$ qui peut être de taille également assez grande (ex. 1000 pour ImageNet). Par ailleurs, on dispose de n échantillons classifiés $\{x_i, y_i = f(x_i)\}_{i \leq n}$ (training set). Le problème sous-jacent ici est l'énorme variabilité à l'intérieur d'une même classe.
- et celui de la **régression**, où la principale différence est que $f(x)$ n'est pas un index mais un réel. La complexité est essentiellement la même car elle vient de la grande dimension d. En Physique traditionnellement pour répondre à une question (ex. répartition de masse dans une galaxie), on étudie les forces fondamentales qui donnent par exemple des équations d'évolution et fournissent l'état du système à instant t par intégration. Cependant, ici nous n'avons pas les équations fondamentales, mais on dispose d'information *a priori* (ex. symétries du système, propriétés de continuité) et la question est de savoir si on peut en déduire par exemple l'énergie du système pour n'importe quelle configuration à partir de quelques configurations connues. On voit ici qu'on a un point de vue totalement différent de la tradition: on ne part pas des interactions fondamentales pour concevoir un modèle et répondre à la question, mais c'est à partir d'exemples qu'on se construit une approximation qui répondra par interpolation à la question posée.

Le problème d'interpolation intuitivement simple est très compliqué pour l'apprentissage. En effet, pour pouvoir faire une interpolation en x il faut pouvoir en quelque sorte avoir des échantillons connus $\{x_i, y_i\}$ dans le voisinage de x pour pratiquer une "moyenne". Par exemple, mettons que les x_i satisfassent la condition suivante (distance euclidienne):

$$\forall x \in [0,1]^d, \exists x_i \in [0,1]^d \quad / \quad ||x - x_i|| \leq \epsilon \tag{1}$$

Si les x_i sont répartis uniformément, alors il faut au moins ϵ^{-d} points qui couvrent tout l'espace $[0,1]^d$. En fait, il faut se rendre compte que les points sont très loin les uns des autres en très grande dimension. Ce thème était celui du cours de 2018, c'est celui de la **malédiction de la dimension**. Moralité pour estimer $f(x)$ en un point $x \in \Omega \subset \mathbb{R}^d$, il faut imposer de **forte régularité de f sur Ω** pour pouvoir l'interpoler entre des points très isolés. La question est de savoir de quelle type de régularité s'agit-il?

Cependant, si les points x_i s'accumulent sur une variété Ω de dimension bien inférieure à d, alors leurs distances seront bien plus petites. Le problème dans ce cas devient "facile", par exemple ceux qui décrivent le mouvement d'un robot articulé, ou celui d'images très simples (ex. classification chiffres binaires). Mais ce type de problème n'est pas celui que l'on considère ici, car prenons le cas d'un image à 10^6 pixels d'une scène de la vie quotidienne, la description nécessite un nombre colossale de variables pour la décrire. Donc, certes x appartient à un sous-ensemble de $\mathbb{R}^d$ mais qui reste de grande dimension et si on l'échantillonne au hasard, on obtient une image de bruit blanc totalement dépourvue de structures. Donc, on a besoin de comprendre la régularité de f mais dans le cadre de l'espace Ω qui définit l'ensemble des images du type qui nous intéresse.

2.3 Changement de variables: représentation

Mettons que l'on dispose d'un lot d'échantillons $\{x_i\}$ avec 2 labels $\{0,1\}$. Ce que l'on aimerait, c'est disposer d'une fonction Φ de $\mathbb{R}^d$ dans $\mathbb{R}^d$ qui transforme les $\{x\}_i$ en $\{x'\}_i$ de telle façon que la frontière entre les deux populations soit un hyperplan:

$$x = (v_1, \ldots, v_d) \xrightarrow{\Phi} \Phi(x) = \{x'\} = (v'_1, \ldots, v'_d) \tag{2}$$

La classification se résume alors en la recherche de l'hyperplan séparateur, c'est-à-dire la recherche du vecteur w normal au plan. Le classificateur a une expression simple (voir

Sec. 1.6 du Cours 2019):

$$\tilde{f}(x) = \text{sign}\left\{\langle w, \Phi(x)\rangle + b\right\} = \text{sign}\left(\sum_k w_k v'_k + b\right) \tag{3}$$

qui revient à faire une combinaison linéaire des coordonnées du vecteur w avec les variables de x dans la représentation donnée par Φ, et à en prendre le signe.

Dans la littérature les éléments $\{\phi^k(x)\}_{k\leq d'} = \Phi(x)$ sont appelés des *features* et l'opération de classification s'apparente à un vote entre des informations faibles du problème pour obtenir une décision plus forte $\tilde{f}(x)$. On a réalisé un changement suivant : la courbe de niveaux (frontière) non-linéaire dans le problème initial a été transformée en une courbe de niveaux linéaire. **Ainsi, on a exprimé la régularité de la fonction $f(x)$ à travers le changement de représentation via Φ.** De plus si ce changement de représentation est simple cela correspond à une fonction f très régulière, et inversement la complexité de la fonction f est vue à travers la dimension de la fonction Φ.

Maintenant, le problème très pratique qui se pose, c'est comment trouver le $\Phi(x)$? Car une fois que l'on a le changement de variable, trouver le vecteur w ce fait simplement en minimisant des critères de marge comme on l'a vu dans les années précédentes: cf. les méthodes SVM, Ridge Regression [8]. Donc, le point dur est celui de trouver Φ. Il y a deux points de vue:

- le premier est assez extrême, il prévalait avant les années 2010 environ et ne repose uniquement que sur l'information *a priori* que l'on a sur la fonction $f(x)$ et on va les coder dans la fonction $\Phi(x)$ de manière à essayer de linéariser le problème.
- le second consiste à apprendre à partir des données: c'est-à-dire que le réseaux de neurones va apprendre le $\Phi(x)$ en même temps qu'il fait la classification/régression.

2.4 Les réseaux de neurones

Un petit rappel [9], ces réseaux de neurones "artificiels" sont introduit dans les années 40 par W. Pitts et W. McCulloch, puis en 1957 F. Rosenblatt en construit un à 1 couche de neurones capable d'apprentissage. Il calcule les coordonnées du vecteur w et un seuil b

8. Voir la Sec. 4.2.4 du Cours de 2019
9. On peut revoir également la Sec. 4 du Cours de 2019 en guise d'introduction.

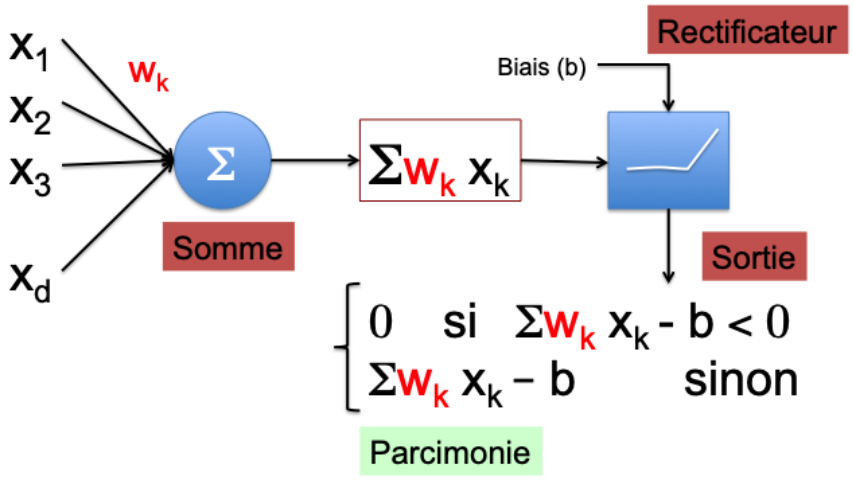

FIGURE 1 – Représentation graphique d'un classificateur linéaire.

qui définissent complètement l'équation d'un hyper-plan séparateur. Sur la figure 1, on a représenté les différentes étapes du Perceptron: les entrées, leur somation pondérée, puis le rectificateur/activateur et enfin la sortie. Notons qu'avec le rectificateur à seuil (type ReLU) la sortie est parcimonieuse puisqu'elle est nulle tant que la somme pondérée est plus petite que ce seuil.

Ensuite, un **réseau à plusieurs couches** est un empilement de Perceptron inter-connectés comme sur la figure 2. La dernière couche consiste à agréger toutes les sorties $\Phi(x)$ pour faire une classification ou une régression. L'apprentissage consiste à optimiser les paramètres inter-couches (poids et biais) pour minimiser l'erreur sur les exemples. Ceci dit c'est un problème difficile d'optimisation, on utilise des méthodes de descente de gradients stochastiques. Le miracle en quelque sorte, c'est que même s'il y a beaucoup de minima, cela marche, à savoir que le réseau est capable de généraliser. Du point de vue de **l'information *a priori***, elle se trouve dans **l'architecture** du réseau et son design est ce qui coûte le plus de temps au bout du compte. Donc, réalisons bien le point suivant: le

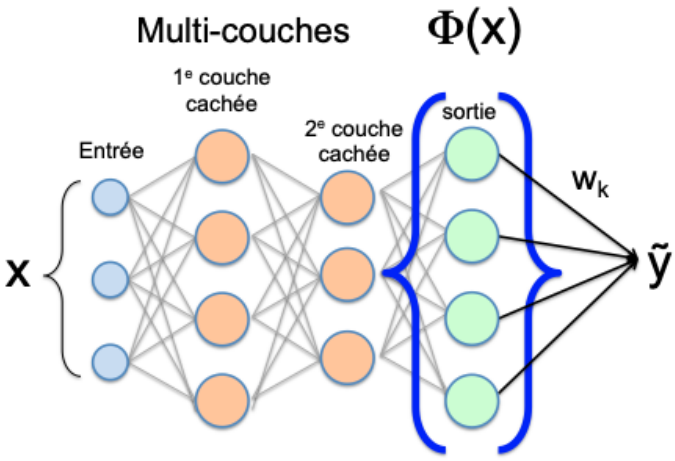

FIGURE 2 – Schéma simplifié d'un réseau de neurones multi-couches où la représentation Φ est le résultat de l'ensemble des sous-couches; elle est utilisée en fin de réseau pour procéder par exemple à une classification.

réseau n'est pas un système qui apprend tout seul, on a introduit un *a priori*.

Une étape très importante dans ce contexte, a été l'introduction des **réseaux convolutionnels** (ou convolutifs, et "convolutional" en anglais) par Y. LeCun et J. Bengio dans les années 1990 [10]. Au centre de ces réseaux, il y a la notion de **filtres** (voir la figure 3) qui vont s'appuyer sur **l'invariance par translation** de beaucoup de problèmes. Qui dit "invariance par translation", dit "opérateur de convolution", d'où la notion de filtre de convolution. Dans la figure 3, cela se traduit par le fait que les poids associés au filtre F1 qui prend en charge des patchs typiquement 3x3, 5x5 sont les mêmes pour tous les patchs de l'image x d'origine. Maintenant, il est d'usage d'empiler différents types de filtres, se sont les F_i. A la couche d'après, il y a un nouvel axe qui apparait, il s'agir de la liste des canaux qui permet de mélanger les différents filtres (figure 4). Donc, l'opérateur qui relie les deux couches (1 et 2) est non seulement invariant par translation mais prenant l'ensemble des petits patchs le long des canaux il faut également fixer les paramètres le long de ce nouvel axe. La grande difficulté est de comprendre la **nature des opérateurs mathématiques** qui agissent le long de cette dimension.

Un autre point important de l'architecture mise au point par Y. LeCun c'est que le

10. voir Sec. 1.8 Cours 2019. Y. LeCun, J. Bengio et G. Hinton ont reçu le Prix Turing 2019.

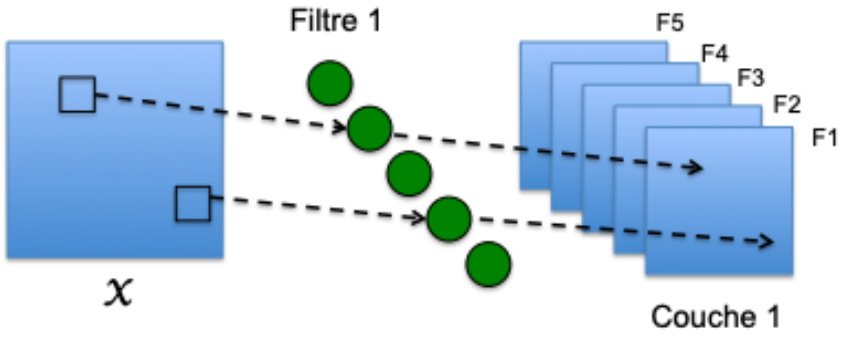

FIGURE 3 – Schématisation d'un premier étage de convolution avec 5 filtres différents. Pour chaque filtre, ex. le filtre F1, chaque neurone s'occupe d'une petite partie de l'image d'origine et tous les poids associés aà chaque neurone sont identiques d'un neurone à l'autre.

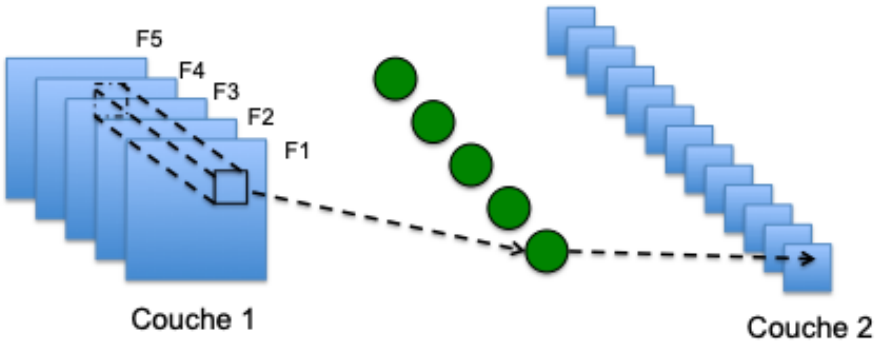

FIGURE 4 – L'opération entre la couche 1 et 2 non seulement agit dans les axes de l'image d'origine mais aussi prend en charge une nouvelle dimension le long des résultats du filtrage précédent. C'est la où réside le point dur de la compréhension de la nature mathématique des opérateurs mis en jeu.

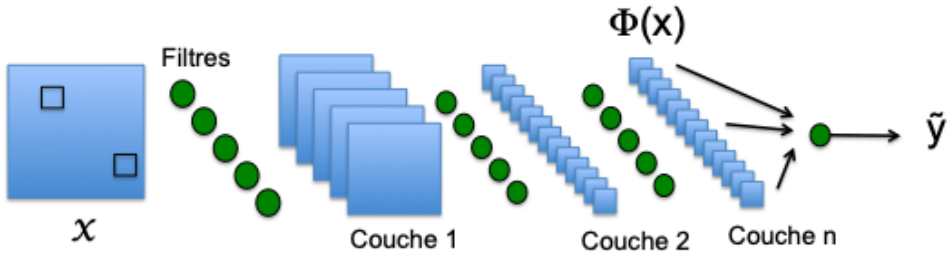

FIGURE 5 – Réseau convolutionel profond constitué de plusieurs couches de filtres, puis en bout de chaine, d'une simple couche linéaire ou de plusieurs complètement connectées qui fournissent la réponse (classification/régression).

support spatial des filtres en entrée (cf. F1) est petit par rapport à la taille de l'image, mais dés la seconde couche il y a un mélange des filtres. Un réseau profond peut se schématiser comme sur la figure 5 qui comporte des millions de paramètres. Donc, il y a beaucoup de structures dans cette architecture qui a été déduite d'une information *a priori*. Ce qui est fascinant, c'est que ce type d'architecture est capable d'attaquer une très grande variété de problèmes: dans le domaine de l'imagerie, du son, du langage, du texte, en physique, chimie etc (voir Introduction du Cours 2019). C'est aspect générique va nous occuper. Notons que $\Phi(x)$ a considérablement moins de paramètres que l'image initiale, c'est-à-dire que la réduction de la dimensionalité est considérable de sorte que la frontière de classification à la fin a été aplatie. Bien entendu la question est de savoir quels sont les **principes de cette architecture**? Il est clair que la notion de **symétrie** est importante, elle vient dès la première couche; de plus il y a la notion de **multi-échelles** car au fur et à mesure que l'on s'avance dans les couches, un neurone "voit" des portions de l'image d'origine de plus en plus grande (cela fait le pendant du système cognitif naturel); enfin il y a la notion de **parcimonie** à travers la réponse neuronale (cf. la fonction d'activation à la ReLU).

Un exemple typique de l'aptitude de ces architectures à faire nettement mieux que les solutions antérieures, est celui de la classification des images de la base ImageNet [11] (voir Sec 2.1.1 du Cours 2019). Avant 2012, les architectures intégraient tous les a priori du problème et procédaient à une régression linéaire alors que les réseaux de neurones

11. http://www.image-net.org/

marchaient moins bien. La donne a changé avec 1) les capacités de calcul accrues, 2) des lots d'entrainement en grand nombre. Ainsi, en 2012 Goeffrey Hinton, Alex Krizhevsky et Ilya Sutsever ont mis au point **AlexNet** qui a supplanté tout le monde. A l'heure actuelle, ImageNet a 1 millions d'images étiquetées selon 2000 classes.

Il est très important encore une fois de répéter qu'il faut beaucoup d'échantillons pour comprendre pourquoi les réseaux de neurones convolutionels profonds fonctionnent si bien, et les comparer à des systèmes qui n'apprennent pas du tout (cf. où les filtres sont fixés à l'avance) ce que l'on abordera dans le cours. Ce qu'on aimerait comprendre bien sûre c'est qu'est-ce qui a été "appris", surtout quand on le compare à un système figé au départ. Notons expérimentalement, par exemple en changeant les conditions initiales des paramètres, la solution de la minimisation a toute les chances d'être différente, et pourtant les réseaux ont les mêmes comportements statistiques (cf. même capacité de généralisation [12]). D'un point de vue mathématique, l'architecture a capturé d'une manière ou d'une autre, une forme de régularité de la fonction qui répond à la question $y = f(x)$.

Pourquoi est-il important de comprendre le "Pourquoi ça marche?" Eh bien, il y a des fois où la généralisation est prise en défaut, et pas qu'un peu! Il s'agit des **exemples adversaires** [13]. Un tel example aussi prés que possible d'un exemple ayant servit à l'entrainement existe, et tous les systèmes ont ce type de pathologies. Il est clair que l'on ne veut pas mettre un tel système dans des appareils où la vie des personnes est impliquées par exemple. Donc, il faut pouvoir disposer d'outils qui garantissent le niveau de sureté d'un système, et donc cela passe par la compréhension de la nature des opérations mathématiques utilisées.

2.5 Systèmes "classiques"

Si on laisse de coté les réseaux de neurones un temps, que sont les systèmes plus "classiques"? Le premier exemple est celui de la reconnaissance de la parole (Voir Sec. 2.2 du Cours 2019) que nous retrouverons dans les cours prochains. **L'analyse temps-fréquence** est un outil de choix, et on se rend compte assez vite de sa grande variabilité (rythme, changement de gamme, etc) selon le locuteur, du spectrograme obtenu qu'en

12. NDJE: on peut même dire même problème face aux exemples adversaires.
13. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I. J., Fergus R., 2013, CoRR, abs/1312.6199.

bien même il s'agit du même mot prononcé. Quelles ont été les techniques utilisées malgré ces variabilités?

Le cas de la parole est intéressant car il a été étudié depuis les années 60, et au fil du temps, il s'est très spécialisé avec une technique qui s'est raffinée, mais qui a très peu évoluée jusque dans les années 2000. L'idée en bref est d'expliciter les structures temps-fréquences représentant les sons, puis on va essayer de reconnaitre les phonèmes qui constituent les mots, et ainsi de suite. Les structures sont les états du systèmes, et il y a des probabilités de transitions d'un état à l'autre (dictées par la connaissance de l'orthographe et de la grammaire). Donc, on définit une chaîne de Markov sur laquelle on y met un modèle de mélange de gaussiennes. Ce type de technique a été complètement transformé par les réseaux de neurones. Tout d'abord, on a changé le modèle des mélanges de gaussiennes, puis ensuite c'est l'analyse complète (cf. comprenant le temps) qui a été prise en charge par les réseaux convolutionnels. Et actuellement, la reconnaissance de la parole est bien mieux traitée par ces réseaux profonds.

Un problème très intéressant est celui de la séparation de sources (dit problème du "Cocktail Challenge") fondamental pour les aides auditives. Il ne faut pas simplement amplifier le tout, car le signal et le bruit ambiant ne sont pas plus séparés. En 2018, une solution a été mis en œuvre par Yi Luo et Nima Mesgarani [14]. La séparation est quasi-parfaite en quelque millisecondes, donc compatible pour des aides auditives. Alors certes le réseau a été entrainé avec beaucoup d'échantillons (qui ne sont pas issus des personnes que l'on doit reconnaitre), mais on ne comprend pas comment cela marche au delà de la première couche de neurones où un semblant de spectrogramme est reconstitué.

Dans le domaine de la Physique (Chimie), il y a l'approche "classique" d'essayer d'intégrer les équations fondamentales (Newton, Boltzmann, Navier-Stockes, Maxwell, Schrödinger...) mais le calcul est difficile dès lors que l'on a un grand nombre d'entités en interaction. L'approche alternative, est de répondre à des questions spécifiques (ex. quelle est l'énergie du système que j'étudie?) à partir d'une base connue d'exemples.

Finalement, quel est le lien entre neurones artificiels et neurones biologiques? Si on arrive à comprendre le fonctionnement des systèmes artificiels lesquels deviennent de plus en plus performants, on est légitimement en droit de se demander que peut-on apprendre sur les structures/fonctionnements des systèmes biologiques. Il y a également un intérêt

14. Voir l'article arXiv:1809.07454v2

pratique à faire l'aller-retour biologique-artificiel, en effet on constate que l'humain en 1/10e de seconde est capable de faire de la reconnaissance d'objets, or les neurones biologiques sont très lents, donc la reconnaissance s'appuie sur très peu de couches (de l'ordre de 7) et qu'il n'y a pas de boucle de rétro-action très complexe. Comment modifier les architectures (artificielles) actuelles pour obtenir ce niveau de performance?

2.6 Questionnement

Donc, la question est : pourquoi les architectures de réseaux convolutionnels sont-elles "génériques ?" Il y a 3 types de problèmes en fait:

- **l'estimation**: analyser l'erreur de généralisation (biais/variance, voir Cours 2018);
- **l'optimisation**: minimiser l'erreur empirique (voir Cours 2019);
- **l'approximation**: lien avec l'architecture, c'est-à-dire avec l'information *a priori* du problème qui est le sujet du cours de cette année.

On va donc se poser les questions: quelle est l'information *a priori*, pourquoi des convolutions, quels types de filtres, quel est le rôle des non-linéarités (cf. activation), et le lien avec la neurophysiologie (contexte de l'image et du son). Nous aborderons les points suivants:

- régularité: séparabilité, symétrie, sparsité/parcimonie
- symétrie: la convolution et l'analyse de Fourier
- séparation et parcimonie: principe d'incertitude et représentations temps-fréquence
- les transformées multi-échelles et ondelettes
- invariance par translations, rotations, déformations
- classification sans apprentissage (SIFT [15] et MFCC [16])
- invariants multi-échelles: réseaux d'ondelettes et scattering
- des applications dans divers domaines
- enfin pourquoi il y a des manques dés que le système devient complexe...

15. Self Invariant Feature Transform
16. Mel Frequency Cepstral Coefficients

3. Séance du (29 Janv.)

3.1 Les questions de base (rappel) et plan du cours

A la séance précédente, nous avons listé un certain nombres de questions:

— Pourquoi les **CNN sont capables de généraliser** sur des problèmes génériques très différents (images, sons, langage, chimie, physique, biologie, etc)? Si tel est le cas, cela veut dire que la fonction à approximer a une régularité particulière.

— Quelle est donc la **régularité générique** (si elle existe) qui serait sous-jacente aux problèmes tels que : la physique, la perception, la biologie, la symbolique (langage)...? H. Simon en 1962 a donné une réponse dans son livre "Architecture of the Complexity" [17]. Cependant, même si ce livre donnait un cadre qui était émergent, ce n'est pas pour autant que l'on était capable d'appréhender les problèmes comme on le fait aujourd'hui. Il y a eu un grand saut effectué entre des arguments qualitatifs et leur mise en pratique.

— S'il y a une régularité générique, c'est qu'il y a de façon concomitante une **information générique** qui est exprimée dans les CNN. Comment l'exprimer mathématiquement (et dans un algorithme)? On sait que $x \in \mathbb{R}^d$ avec d très grand, mais x a de la structure. En particulier, x est indexé selon $x(u)$ (cf. u un index d'un pixel dans une image, une trame temporelle de sons, un mot dans un texte, etc). Or, u est dans **un espace à basse dimension** (ex. $u \in \mathbb{R}$ pour un son, $u \in \mathbb{R}^2$ pour une image, $u \in \mathbb{R}^3$ pour une vidéo, etc), et donc on a une structure en basse dimension qui va donner essentiellement le contenu de **l'information *a priori*** que l'on va pouvoir "hard coder" dans l'architecture du réseau. En effet, **on peut à travers u imposer des symétries du problème**: ex. la translation donc toute l'analyse de Fourier va en découler, plus généralement les structures de groupe qui vont agir sur la variable u, les voisinages donc les représentations multi-échelles, on peut également utiliser des graphes pour rendre compte des structures d'interaction entre des variables temporelles (ex. prix d'actions en Finance, des acteurs dans les réseaux sociaux, etc). Donc, **c'est à travers les structures à basses dimensions que les mathématiques actuelles ont un angle d'attaque du problème**.

Le plan du cours de cette année 2020 est le suivant:

17. voir Sec. 4.1 Cours 2019 sur "La cybernétique".

1. Architecture de la Complexité, Estimation, Optimisation, Approximation
2. Approximation: Multi-échelle, Groupe et Symétrie (réduction de dimensionalité)
3. Temps/fréquences, principe d'incertitude, localité (problématique qui va au delà des CNN)
4. Échelle : Transformée d'Ondelettes (TO), Perception auditive
5. Ondelettes dyadique Banc de Filtres (lien avec les invariants)
6. Perception visuelle, (TO et invariants) et formes non-linéaires (Scattering)
7. Applications: Image/son, chimie quantique. Limitation de l'introduction de l'information *a priori* qui n'est pas suffisante pour expliquer les performances des CNN.

Une remarque, on a dit que par l'étude des CNN on allait revoir l'analyse harmonique classique. En cela les CNN vont donner un éclairage différent, en particulier parce qu'ils sont non-linéaires alors que l'analyse harmonique est essentiellement linéaire. Ainsi, il faudra s'interroger sur le rôle de ces **non-linéarités**.

3.2 Architecture de la Complexité

En 1962, on est dans le domaine de la "Cybernétique", introduite par le mathématicien N. Wigner (1947) [18], qui a pour leitmotiv d'analyser l'évolution d'un système complexe par une boucle de rétroaction (Fig. 6). L'idée maitresse est que l'on s'abstient de modéliser le système complexe (la "boîte noire"), ce qui importe est le but (réponse) recherché. On peut par ce schéma assez simple, comprendre le comportement de systèmes très complexes d'interactions entre agents par exemple, également on peut appréhender un des grands systèmes bouclés qu'est celui de l'évolution.

Maintenant, si on essaye d'ouvrir la "boîte noire" pour comprendre la structure interne, H. Simon fait ressortir plusieurs idées fondamentales:

— la structure est **hiérarchique** quasiment tout le temps;

— une **explication dynamique** (temporelle) de cette structure hiérarchique est la recherche de stabilité (survie);

18. Quelques domaines dans lesquels Wigner a eu un impact décisif: les probabilité avec la mesure de Wigner, l'analyse harmonique revue, le traitement du signal et du contrôle.

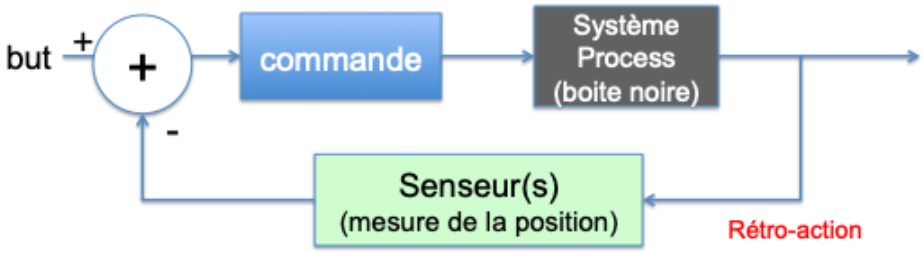

FIGURE 6 – Le principe de l'analyse de l'évolution d'un système complexe par boucle de rétroaction.

— **séparabilité des échelles** (dans la hiérarchie) qui permet de détruire la malédiction de la dimension;
— la description temporelle doit être vue comme des **processus** et non une succession d'états statiques. Ceci fait résonance avec la notion d'apprentissage actuel des CNN où les données d'entrainement sont analysées séparément sans lien entre elles, ce qui expliquerait la nécessité d'en disposer en grand nombre, alors que l'humain est capable d'apprentissage à moindre coût car les images s'inscrivent dans une dynamique. Il est clair que la base de la dynamique (ex. physique) ce sont les équations différentielles.

3.2.1 Structures hiérarchiques

Le constat de l'existence de structures hiérarchiques est manifeste: en Physique, on a beaucoup d'exemples de structures et de hiérarchies depuis les particules élémentaires (micro) jusqu'aux matériaux (macro), idem en Biologie où l'on peut partir des molécules (micro) jusqu'aux organismes complexes (macro); mais aussi en Perception des pixels aux scènes en passant par des objets; les Systèmes symboliques de la lettre aux livres; comme enfin dans les Systèmes Sociaux. Les contres exemples où il n'y aurait pas de structures hiérarchiques présentent par contre des symétries notables.

Le Pourquoi de l'existence de ces structures hiérarchiques est la question qui vient immédiatement. H. Simon donne une réponse qui est liée à **l'évolution dynamique du**

système. Selon lui, pour qu'il y est évolution, il faut que tous les états intermédiaires (cf. sous-structures) soient stables [19]. Le nombre de sous-structures avec lesquelles il y a une interaction directe est assez limité, et les interactions fortes sont à l'intérieur des structures, alors que les interactions entre structures sont plus faibles. Ainsi, la dynamique haute fréquence est gouvernée par l'intérieur des structures, et la dynamique basse fréquence est gouvernée par les interactions inter-structures.

La grande difficulté est que la description sous forme d'arbre qui tente à capturer la structuration hiérarchique (verticale) ne marche pas la plus part du temps, car il y a de la structuration horizontale et il y a de fait une interaction avec toutes les structures au bout du compte. Il y a donc une **séparabilité faible** qui complique la mise en forme d'une solution par arbre. Il y a un autre point qui est également fondamental est la notion de **symétries** liées à ces structures. Par exemple, dans le monde micro, la fonction d'onde a des formes particulières engendrées par l'indiscernabilité des électrons qui la compose; dans le monde macro la symétrie de l'architecture de certains bâtiments est manifeste (cf. par exemple les fenêtres sont interchangeables sans que pour autant l'édifice en pâtisse); idem pour le cas de personnes au sein d'une entreprise où pour un certain nombre de taches les personnes sont interchangeables, etc. Il est clair que plus il y a de symétries moins il y a besoin de structures élémentaires: pensons à la brique de LEGO qui permet une foultitude de structures complexes, en biologie on ne compte qu'une vingtaine d'amino-acides, le nombre d'atomes dans la table de Mendeleïev est de l'ordre de 94 si on se restreint à ceux rencontrés dans la Nature,... Donc il y a une forme de **parcimonie** sous-jacente. Mais, si le nombre de structures élémentaires est relativement petit, il n'en reste pas moins que lorsque l'on empile les structures cela devient plus compliqué: l'interchangeabilité entre humains a des limites qu'en même.

3.2.2 La description temporelle

En schématisant, la description temporelle peut s'aborder par deux approches:

— une première par **l'état** $x(t)$ dont le temps sert d'indice, et d'un instant à l'autre l'état change (ex. un diaporama de photos de famille); le problème est que si on veut avoir une description détaillée il faut pouvoir disposer d'un très grand nombre d'états.

19. Voir Sec. 4.1 Cours 2019, la métaphore de l'horloger.

— l'autre par **évolution du processus** qui met en jeu des équations différentielles de type

$$\frac{\partial x(t)}{\partial t} = H(t)x(t)$$

avec l'hamiltonien $H(t)$, l'opérateur d'évolution. Et donc en principe on devrait s'attacher à l'étude de cet opérateur d'évolution. Or, ce n'est pas ce qui est fait dans les réseaux de neurones (type CNN). Ceci dit dans le cours de cette année, nous étudierons des problèmes sans temps, c'est-à-dire des problèmes de type $x(u)$ bien que certainement cette approche est très sous-optimale. Par exemple, quand on observe le système de perception visuel du cerveau, on constate que les boucles de rétro-action ont des constantes de temps bien plus grandes que celles des processus feed-forward. Par exemple, la reconnaissance d'une image (visage) se fait sur 1/10sec sans boucle de rétro-action, c'est-à-dire comme pour une analyse statique, ce qui a amené les développements en oubliant le temps. C'est-à-dire que pour reconnaitre une image il n'y a pas besoin de processer une vidéo. Par contre, la reconnaissance d'un très grand nombre d'images, sans vidéo il faut pouvoir disposer d'un nombre colossal d'images pour faire l'entraînement. Donc, en résumé, il est clair que pour étudier certains types de problèmes, il n'y aurait pas lieu de considérer le temps, cependant on le paye par l'inefficacité à l'entrainement.

3.3 Les réseaux CNN

D'une certaine façon, les réseaux de neurones sont des systèmes qui tentent de répondre au programme d'H. Simon: l'apprentissage d'un réseau de neurones est un système à boucle fermée, où la fonction de coût joue le rôle de la mesure de l'erreur qui influe par réaction sur l'ajustement des paramètres. Le gros avantage est que l'on dispose d'un algorithme que l'on peut analyser.

Donc, l'entrée [20] $x(u)$ va être transformée par la première série de filtres en $x_1(u,k)$ où k est un indice de canal, l'opération étant une convolution, puis il y a une non-linéarité ρ de type ReLU (Rectified Linear Unit) [21]. Donc, la relation de passage de x à x_1 est donnée

20. Notons que si l'on considère une image, celle-ci peut être elle-même constituée de plusieurs canaux, ex. RGB quand on prend une photo en couleur, ou même *ugriz* si l'on considère des filtres astronomiques.
21. $\rho(x) = \max(0, x)$

par :

$$x_1(u,k) = \rho(x * h_k(u)) \tag{4}$$

Ensuite, la plus part du temps s'en suit une opération de sous-échantillonnage (ex *max-pooling* ou *average-pooling*) .

La couche suivante est plus compliquée, car un filtre est appliqué non plus sur 1 patch d'une image mais sur un cube de données. Ensuite on répète les couches de convolution/pooling pour aboutir à la représentation $\Phi(x)$ qui dépend de tous les opérateurs qui définissent les filtres utilisés, à savoir les paramètres du système sont $\theta = \{L_i\}_{i \leq p}$ pour un système à p couches [22]. Donc, avec un réseau de neurones à p couches, on peut approximer une classe de fonctions, définie par l'ensemble $\mathcal{H}$ tel que:

$$\mathcal{H} = \{f_\theta \ / \ \theta = \{L_i\}_{i \leq p}\} \tag{5}$$

Afin de déterminer les paramètres θ adaptés au problème posé, on va minimiser l'erreur entre le "vrai" f et le f_θ. On se rend compte immédiatement, si le nombre de paramètres est de plusieurs millions voir milliards, que $\mathcal{H}$ est de dimension colossale. La complexité des modèles est donnée par $\log |\mathcal{H}|$.

Maintenant, afin de comprendre comment un réseau généralise, il faut adresser les trois domaines suivants: **l'estimation** (statistique) qui va permettre d'obtenir θ à partir d'une base d'apprentissage $\{x_i, y_i = f(x_i)\}_{i \leq n}$; **l'optimisation** qui consiste à minimiser l'erreur par exemple à travers une méthode de descente de gradients stochastiques; enfin il y a le problème **d'approximation** qui va consister à regarder l'erreur minimum en choisissant le meilleur θ, puis à savoir si l'erreur décroit quand on augmente la taille de θ (cf. le cardinal de $\mathcal{H}$). Remarquons, avant de plonger dans le détail, que ces trois domaines sont imbriqués alors que la communauté en Machine Learning est assez scindée entre les statisticiens qui vont s'attacher à l'estimation, les optimiseurs qui vont œuvrer pour rendre le problème numérique le plus efficace/robuste, enfin il y a des spécialistes de la théorie de l'approximation qui réponde à la question de la taille de $\mathcal{H}$ pour pouvoir approximer f avec avec une erreur ϵ. Ces trois communautés disjointes ont été contraintes, si on peut dire, à évoluer pour attaquer le problème des réseaux, en particulier l'estimation et

22. nb. dans le cas d'un ReLU classique, il n'y a pas de paramètre associé, ce qui n'est pas le cas par exemple d'un PReLU qui donne une valeur légèrement a non-nulle pour une entrée négative, laquelle est optimisée en même temps que les paramètres des filtres.

l'approximation sont devenus des problèmes très proches l'un de l'autre.

3.4 L'estimation: biais-variance

C'est un sujet qui a été abordé durant le cours de 2018, et on en donne ici les idées essentielles. L'enjeu est de pouvoir approximer f (inconnue) à l'aide de f_θ. Donc, on se donne un risque (erreur) R définit par l'espérance de l'erreur entre la prédiction $f_\theta(x)$ et la vraie valeur y, selon

$$R(f_\theta) = \underset{(x,y)\sim\Omega}{E}[r(f_\theta(x), y)] \tag{6}$$

où Ω est l'espace de distribution jointe qui relie x à y. On veut minimiser $R(f_\theta)$ et donc trouver:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\ R(f_\theta) \tag{7}$$

Selon le problème (régression/classification) la nature de la fonction r change. Par exemple on utilise souvent un risque quadratique pour $y \in \mathbb{R}$. Dans le cas d'une classification à K classes, le réseau sort un tableau z à K valeurs indexées par y. L'idée est que la réponse soit d'autant plus grande quelle pointe sur la bonne classe. Pour se faire on va donner une estimation de la probabilité jointe de y et de x indexée par θ, soit $p_\theta(x, y)$. Ce faisant, on se place dans un cadre d'estimation probabiliste (alors que nous étions dans un problème d'approximation) qui nous conduit à la notion de **maximum de vraisemblance** où θ va maximiser $p_\theta(x_i, y_i)$ **sur les exemples de la base d'entrainement**, ce qui revient à minimiser $-\log p_\theta(x_i, y_i)$.

Cependant, la sortie du réseau n'est pas une distribution de probabilité (cf. la somme n'est pas égale à 1), et donc pour se faire on utilise la fonction `softmax` (les sommes sur y courent sur les K composantes du vecteur de sortie, et dans ce cas la "cible" est codifiée dans un hot-vecteur dont la seule composante non-nulle est celle de la bonne classe):

$$f_\theta(x) = z_y(x) \xrightarrow{\text{softmax}} \frac{e^{z_y(x)}}{\sum_{y'} e^{z_{y'}(x)}} = p_\theta(x, y) \tag{8}$$

Donc, en sommant sur l'ensemble des échantillons, la fonction à minimiser s'écrit

(notons que θ est inclus dans le calcul de $z_y(x)$):

$$L(\theta) = \sum_i -\log\left(\frac{e^{z_{y_i}(x_i)}}{\sum_{y'} e^{z_{y'}(x_i)}}\right) = -\sum_i \left[z_{y_i}(x_i) - \log\left(\sum_{y'} e^{z_{y'}(x_i)}\right)\right] \tag{9}$$

Si, $z_{y_i}(x_i)$ réalise le maximum, alors la fonction est minimum, et l'avantage est que $L(\theta)$ est différentiable ce qui rend possible l'utilisation d'un algorithme de descente de gradient. Ainsi, le problème d'approximation/estimation tient compte de l'usage d'un algorithme d'optimisation.

Donc, si on revient à notre problème initial d'estimation de $R(f_\theta)$, finalement on est uniquement capable de manipuler un risque empirique $\tilde{R}(f_\theta)$ obtenu à partir des exemples de la base d'entraînement, c'est-à-dire

$$\tilde{R}(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} r(f_\theta(x_i), y_i) \tag{10}$$

Donc, on estime l'espérance à l'aide d'une moyenne empirique. Soit $\tilde{\theta}$ la valeur des paramètres qui minimise le risque empirique

$$\tilde{\theta} = \underset{\theta}{\operatorname{argmin}}\ \tilde{R}(f_\theta) \tag{11}$$

Il faudrait être capable de contrôler la différence (l'erreur) entre θ^* et $\tilde{\theta}$, ou bien les risques correspondants.

On va regarder comment se comparent le (vrai) risque minimum $R(f_{\theta^*})$ et le (vrai) risque de généralisation donné par $R(f_{\tilde{\theta}})$ qui est donné par exemple en donnant au réseau des lots de test.

Prop. [23]

$$\boxed{R(f_{\theta^*}) \leq R(f_{\tilde{\theta}}) \leq R(f_{\theta^*}) + 2\ \max_{\theta} |R(f_\theta) - \tilde{R}(f_\theta)|} \tag{12}$$

23. Voir aussi Sec. 2.3 du cours de 2018.

L'inégalité de gauche va de soit par définition de $R(f_{\theta^*})$. Pour l'inégalité de droite

$$\begin{aligned} R(f_{\tilde{\theta}}) - R(f_{\theta^*}) &= R(f_{\tilde{\theta}}) - \tilde{R}(f_{\tilde{\theta}}) + \tilde{R}(f_{\tilde{\theta}}) - \tilde{R}(f_{\theta^*}) + \tilde{R}(f_{\theta^*}) - R(f_{\theta^*}) \quad (13) \\ &\leq 2 \max_{\theta} |R(f_\theta) - \tilde{R}(f_\theta)| + \tilde{R}(f_{\tilde{\theta}}) - \tilde{R}(f_{\theta^*}) \quad (14) \end{aligned}$$

or $\tilde{R}(f_{\tilde{\theta}})$ est minimum quand on considère le risque empirique $\tilde{R}$, donc $\tilde{R}(f_{\tilde{\theta}}) - \tilde{R}(f_{\theta^*}) \leq 0$ et ainsi

$$R(f_{\tilde{\theta}}) - R(f_{\theta^*}) \leq 2 \max_{\theta} |R(f_\theta) - \tilde{R}(f_\theta)| \quad (15)$$

Donc, quand on remplace le paramètre optimal θ^* par le paramètre appris $\tilde{\theta}$, le risque est plus grand (première inégalité), mais l'erreur est gouverné par **l'erreur que l'on fait quand on approxime le risque vrai par le risque empirique** (seconde inégalité).

Cette erreur dépend donc de la qualité d'approximation de la fonction. On se retrouve alors dans le contexte du problème de statistique de la concentration d'un estimateur autour de sa moyenne. L'erreur sur l'estimation va dépendre de 2 termes: un terme de **biais** irréductible donné par $R(f_{\theta^*})$ qui constitue un minimum, et un terme de **variance** due à la fluctuation de l'estimation sur θ.

En apprentissage, on va utiliser une hypothèse cruciale qui n'est pourtant pas toujours valide, à savoir que les observations d'apprentissage sont **indépendantes**. Ceci est fondamental, car pour qu'une moyenne converge vers une espérance; il faut que les fluctuations se compensent ce qui est beaucoup plus favorable si les échantillons sont iid. C'est un des problèmes: **il faut s'assurer que les échantillons ne soient pas biaisés et qu'ils suivent bien la distribution "générale" du problème**. Moyennant ces hypothèses, et la condition que la fonction r est suffisamment régulière, on peut démontrer (cf. Théorème PAC du Cours de 2018):

Théorème 1. *Avec une probabilité $P \geq 1 - \delta$, on a*

$$\boxed{\max_{\theta} |R(f_\theta) - \tilde{R}(f_\theta)| \leq \sqrt{\frac{\log |\mathcal{H}| + \log(2/\delta)}{n}}} \quad (16)$$

Le terme en $1/\sqrt{n}$ est classique quand on prend n échantillons indépendants, et le numérateur se compose d'un terme lié à la taille de la classe d'hypothèses et l'autre au

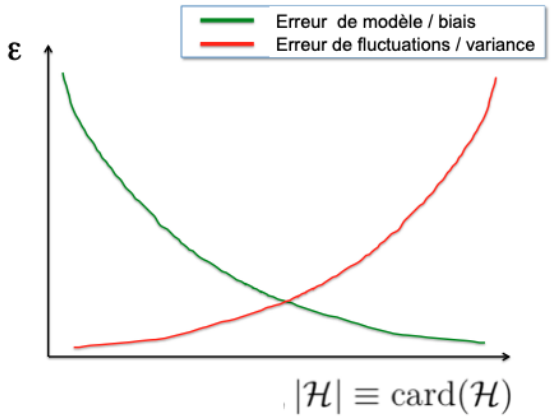

FIGURE 7 – Erreur de Biais-Variance en fonction de la taille de l'ensemble $\mathcal{H}$.

niveau de confiance que l'on veut atteindre. D'un coté plus δ est petit, plus il va falloir d'échantillons (n grand), de l'autre à n et δ fixé on doit choisir la classe d'hypothèses judicieusement pour que $\log|\mathcal{H}|$ (cf. le nombre de paramètres) ne soit pas trop grand. Ce résultat est le point de vue classique de l'estimation de l'erreur.

La conséquence pratique est qu'il y a un système de vase communiquant entre le biais et la variance quand on regarde l'évolution de l'erreur en fonction de la taille de $\mathcal{H}$ (voir Fig. 7). Donc, il faut ni un modèle pas trop grand sous peine d'overfittng, ni un modèle trop petit qui underfit.

Comment donc limiter la taille de $\mathcal{H}$ qui revient à limiter la zone d'exploration des valeurs de θ? Cela peut se faire à l'aide d'une **régularisation**. Un premier type de régularisation va pénaliser la loss par exemple par une norme de type $\|\theta\|^2$ ([24]). Un second type consiste à faire un arrêt de l'optimisation (*early stopping*).

Ceci dit, la nécessiter de limiter le nombre de paramètres n'est pas vraiment ce que constate ceux qui pratiquent les réseaux de neurones. Au contraire, on tend à montrer le contraire à savoir que cela marche d'autant mieux que l'on utilise des réseaux de plus en plus grands. Récemment on à observer, contrairement à la courbe totale des 2 termes biais/variance, plutôt une courbe du type de la figure 8. C'est-à-dire que lorsque le nombre

24. NDJE: cf. Sec. 4.2.3 du Cours 2019, et voir également Sec. 7.2.4.2 qui donne le point de vue bayésien et fréquentiste sur le sujet. Notons aussi que la technique du *weight decay* permet d'introduire une régularisation en L2.

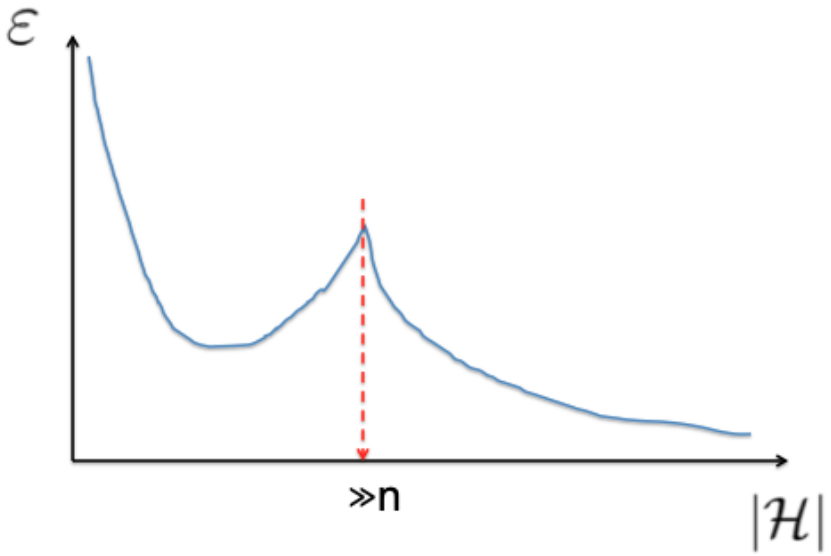

FIGURE 8 – La borne supérieure classique est battue dans les réseaux de neurones très profond, au delà d'une taille critique, l'erreur diminue au lieu de s'envoler.

de paramètres dépassent une taille critique, l'erreur (ici de généralisation) recommence à décroitre. Ce qui veut dire que l'on a été capable de battre la borne supérieure "classique" du théorème PAC.

Comment expliquer ce phénomène? On commence à avoir de la redondance dans les θ, c'est-à-dire que dans l'espace des paramètres on augmente semble-t'il la densité d'échantillonnage. **La transition se fait quand le nombre de paramètres atteint une valeur critique bien supérieure au nombre d'exemples.** Cependant, ces courbes de "double-slopes" qui sont issues de résultats récents ne semblent pas expliquer les résultats actuels sur ImageNet car on se situerait du coté gauche de la courbe. Ce qui est important est que dès que les paramètres ne sont pas indépendants, on observe des phénomènes collectifs.

3.5 L'Optimisation

Comment obtenir $\tilde{\theta}$? On procède à une descente de gradient stochastique par batch [25] et l'on utilise l'algorithme *back-prop*. Mais, le problème est fortement non-convexe, donc il est facile de se faire piéger dans des minima locaux. Cependant, on constate que malgré les conditions de minimisation différentes qui aboutissent à des solutions différentes, les propriétés de généralisation sont identiques. Face à ce constat, les mathématiques vont tenter de décrire le paysage de la fonction de coût lors de la minimisation, et de démontrer que les minima locaux "loin" du minima absolu vont être rares et extrêmement étroits, donc que leurs bassins d'attraction sont peu probables et que donc finalement ils ne sont pas gênants [26]. Ce type de problème est identique à celui que l'on rencontre en physique statistique quand on cherche à minimiser une fonction d'énergie. A coté de ce problème, en apprentissage, on va ajouter des formes de régularisation: *early stopping*, pénalisation L2, ou drop-out.

Ceci étant dit, ce qui va nous intéresser par la suite c'est le problème de **l'approximation** qui correspond à la courbe de biais de modèle (Fig. 7), c'est-à-dire comment configurer le réseau pour que l'erreur obtenue décroisse rapidement en fonction du nombre de paramètres.

4. Séance du (5 Février.)

4.1 Estimation/Optimisation

On commence par un rappel de quelques notions pour fixer les idées. Nous avons donc à comprendre la classe $\mathcal{H}$ des fonctions f_θ que peut approximer un réseau. Rappelons que θ est l'ensemble des paramètres du réseaux (cf. des filtres), et notre objectif est de minimiser le risque en moyenne

$$R(f_\theta) = \underset{(x,y)\sim\Omega}{E}[r(f_\theta(x), y)] \tag{17}$$

25. NDJE: Voir Sec. 9 Cours 2019
26. NDJE: quid des points scelles?

Idéalement, on aimerait trouver l'algorithme qui trouve θ^* qui minimise ce risque "vrai". Or, ce qui (pour le moment) nous est accessible c'est plutôt un risque empirique calculé avec un lot d'entrainement:

$$\tilde{R}(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} r(f_\theta(x_i), y_i) \tag{18}$$

dont la minimisation peut nous donner $\tilde{\theta}$. Ainsi, le risque de généralisation $R(f_{\tilde{\theta}})$ est borné par l'équation 12. **L'erreur d'approximation, aussi appelé "biais de modèle"**, est donnée par la courbe (verte) de la figure 7([27]) qui **décroit en fonction de la complexité du réseau** (cf. le cardinal de $\mathcal{H}$). L'autre contribution due aux fluctuations de la fonction f_θ dans la classe $\mathcal{H}$ est donnée par la courbe croissante (rouge) en fonction de $|\mathcal{H}|$, elle est bornée par l'expression (Eq. 16) du théorème PAC. La taille "idéale" de la classe $\mathcal{H}$ est typiquement celle pour laquelle les deux contributions sont égales.

Donc, on en déduit immédiatement que l'on n'a pas intérêt à se doter d'une classe trop grande, car il y aurait trop de paramètres (cf. le $\log|\mathcal{H}|$ augmente) qui nécessiterait d'autant plus d'échantillons (cf. n). Or, ce n'est pas ce que l'on constate quand on manipule les réseaux de neurones: à partir d'un seuil critique, plus le réseau est grand, meilleur est son pouvoir de généralisation. **Ce constat est en complète contradiction avec la prédiction précédente.** D'ailleurs, il est en contradiction avec l'intuition que l'on pourrait formuler à partir d'un problème d'interpolation. La question est Pourquoi?

Mikhail Belkina et collaborateurs donnent une description du phénomène dans un article récent [28] . L'observation est la suivante (Figure 9): à coté de la courbe classique de l'erreur d'approximation donnée par $R(f_{\theta^*})$, il faut considérer **la courbe du risque empirique** (erreur d'entrainement) donnée par $\tilde{R}(f_{\tilde{\theta}})$ qui pour les grands réseaux atteint 0 au point appelé: **interpolation threshold**. C'est-à-dire, si on prend l'image de l'interpolation 1D, on a trouvé une fonction qui "passe" par toutes les données exactement. Mais on devrait s'attendre à ce que le réseau ne puisse pas généraliser pour des exemples qui ne sont pas dans la base d'entrainement, or ce qui est constaté c'est qu'à partir de ce point il y a une nouvelle décroissance de l'erreur totale pour des réseaux encore plus grands.

27. NDJE: Attention dans des versions de mes notes sur le cours de 2018, il y a une inversion dans la couleur de la légende qui peut prêter à confusion.

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

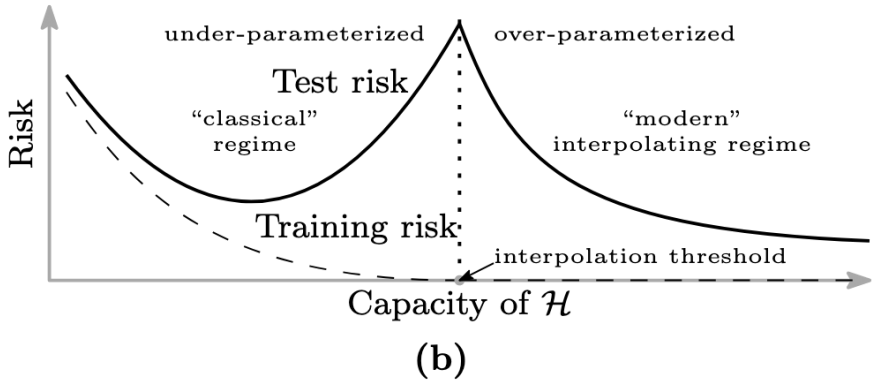

FIGURE 9 – Extrait de la figure 1 de l'article arXiv:1812.11118v2

La courbe de l'erreur totale est appelée la **"double descent risk curve"**. En pratique, on a donc le message suivant: augmenter la taille du réseau franchement pour battre le minimum de la courbe "classique" qui a lieu avant l'interpolation threshold.

Peut-on comprendre cette 2eme descente du risque? **On est clairement au delà du domaine de l'interpolation, le nombre de paramètres dépasse largement le nombre de degrés de liberté**. On a une forme de redondance sur les paramètres, et la pente de la seconde descente dépend de l'algorithme d'optimisation. Les SGD (stochastic gradient descent) ont intrinsèquement des mécanismes de régularisation grâce au bruit aléatoire sur les gradients qui fait que parmi toutes les solutions (sur θ), l'algorithme converge vers des formes les plus régulières possibles. Donc, le comportement de cette seconde descente (sujet d'étude actuelle) est lié à un phénomène de régularisation, et toute la question est de savoir si on a vraiment intérêt à faire de l'hyper-paramétrisation et laisser la main à l'algorithme de minimisation, ou bien s'il n'est pas plutôt préférable de procéder à une régularisation explicite et direct du risque (voir la figure 10)? et la question qui suit est: de quel type de régularisation s'agirait-il d'utiliser? Ce type de questions est très actuel et les réponses dépendent des cas particuliers. On en retient qu'il faut beaucoup de paramètres, cependant savoir si on se trouve avant ou après *l'interpolation point*, dépend du nombre d'exemples d'entrainement et de la complexité du problème. Par exemple avec ImageNet, on n'a pas atteint le point d'interpolation, alors que pour MNIST avec des réseaux de taille raisonnable, on peut aller delà du point d'interpolation et observer la seconde descente.

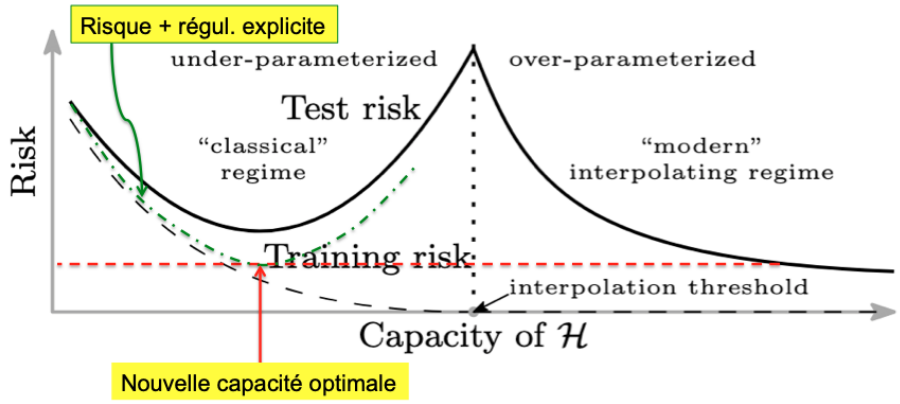

FIGURE 10 – En reprenant la figure 9 sans aller dans la région d'hyper-paramétrisation, est-il possible de régulariser explicitement le risque pour obtenir la courbe vert point-tirer afin d'obtenir le même minimum mais pour un nombre de paramètres bien moindre, ce qui permettrait de faciliter l'entrainement.

Ainsi, on prend pleinement conscience que le problème d'estimation est totalement combiné au problème d'optimisation, il faut penser les deux en même temps. La question qui vient donc à l'esprit est la suivante: si on a bien optimisé, qu'en est-il du minimum de l'erreur? C'est-à-dire quelle est la vitesse de convergence asymptotique? En effet, plus la décroissance de $R(f_{\theta^*})$ est rapide, moins l'erreur de fluctuation sera importante. L'idée est bien qu'en même de pouvoir répondre au problème en utilisant le moins de paramètres possible.

4.2 Le problème d'approximation

En 2018, nous avons étudié le **problème de la malédiction de la dimensionalité**, et nous faisons ici qu'un bref rappel des principales idées. Le point central vient de ce que x (échantillon/donnée) est élément de $\mathbb{R}^d$ avec d très grand et donc il faut imposer une forme de régularité de la fonction que l'on utilise afin de limité la classe de $\mathcal{H}$. De quelle régularité parle-t'on? par exemple la continuité (régularité faible), dérivabilité (régularité forte), etc. On a vu également une régularité dite localement lipschienne (régularité intermédiaire) et uniformément lipschienne sur un espace que l'on peut considérer comme des formes de

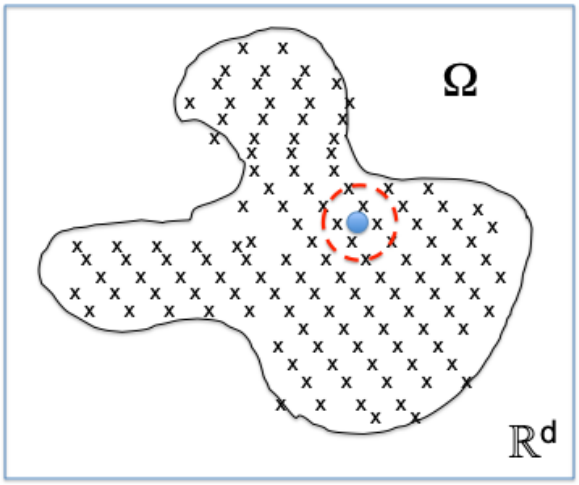

FIGURE 11 – Pavage de l'espace Ω à l'aide de boule de rayon ε.

dérivabilité [29]:

— **Définition**: f est **uniformément Lipschitz** sur $\Omega \subset \mathbb{R}^d$ si

$$\exists\, C/\ \forall (x, x') \in \Omega \quad |f(x) - f(x')| \leq C||x - x'||$$

Que peut-on en déduire sur la capacité à faire de l'approximation avec ce type de régularité? Ayant des échantillons x_i pour lesquels on connait $y_i = f(x_i)$, on veut pouvoir donner une approximation de $f(x)$. Donc, il faut étudier la distance $||x - x_i||_{i \leq n}$ et l'on voudrait que le minimum de cette distance soit inférieur à ε. Cependant, de combien de boules de rayon ε faut-il pour paver tout l'espace Ω (figure 11)?

Prop. [30]: Si $\Omega = [0, 1]^d$ le rayon ε de n boules qui recouvrent Ω satisfait la relation

$$\varepsilon \approx \sqrt{d}\, n^{-1/d} \tag{19}$$

donc pour satisfaire le critère, il nous faut pouvoir disposer d'un nombre d'échantillons dont la loi d'échelle est

$$n \geq C\varepsilon^{-d} d^{d/2} \tag{20}$$

Avec $d \sim 30$, on voit clairement que ce nombre est colossal, même en relativement "basse"

29. NDJE: On pourrait également introduire la notion de fonctions höldériennes.
30. NDJE: voir cours 2018 sec. 3.4; il y a une constante dans la borne sur ε qui est donné par $1/\sqrt{2\pi e}$).

dimension. On conclut également que

$$\boxed{\|f - f_\theta\|_\infty = \sup_{x\in\Omega}|f(x) - f_\theta(x)| \leq C\sqrt{d}\; n^{-1/d}} \tag{21}$$

c'est-à-dire qu'il y a bien une décroissance en fonction de n, mais elle est exponentiellement lente. A partir de là deux réflexions peuvent être élaborées:

— Quid de Ω? sa dimensionalité est peut-être beaucoup plus petite que d. Certes, il y a des problèmes simples de système robotique à petit nombre de degrés de libertés, cependant dès que l'on s'attaque à des images, des sons, des phénomènes sociaux même si la dimensionalité de Ω est peut-être plus petite que d, elle n'est "pas si petite que cela" pour que le problème de la malédiction de la dimension ne soit pas loin.

— Prenons pour acquis que la dimension de Ω est grande, alors **il faut imposer une régularité plus forte** que celle des fonctions lipschiziennes. Il faut comprendre en fait que l'on ne peut adopter des contraintes locales, car la probabilité de trouver "suffisamment" d'échantillons pour obtenir une interpolation au point x d'intérêt est quasi nulle. Donc, il faut se tourner vers des **régularités fortes globales**. Tout le problème est de trouver/définir ces régularités uniformes. C'est là où le texte d'H. Simon permet de donner des pistes d'investigations en analysant les types de hiérarchies et interactions entre "acteurs" de systèmes complexes.

4.3 Régularités globales: séparabilité, symétrie

Tout d'abord faisons un petit retour sur les notions de **séparabilité, symétrie et parcimonie** que nous analyserons en détails sur des cas concrets par la suite.

4.3.1 Séparabilité des variables

C'est la plus forte des notions en quelque sorte: on espère que le problème soit **séparable en basse dimension**. C'est une hypothèse qui vient de la réflexion suivante: $f(x)$ est une fonction en principe de d variables, mais supposons que le problème permette l'écriture suivante:

$$f(x) = f_1(P_{V_1}x) + f_2(P_{V_2}x) + \cdots + f_K(P_{V_K}x) \tag{22}$$

où les P_{V_k} sont des projections orthogonales sur des ensembles tels que $\dim(V_k) \leq q$. $P_{V_k}x$ a q variables (au plus) qui sont des combinaisons linéaires des d variables originales. Un cas particulier:

$$f(x_1, \ldots, x_d) = \sum_{k=1}^{K} f_k(x_{i_1^k}, \ldots, x_{i_q^k}) \tag{23}$$

c'est-à-dire que les fonctions f_k sont des fonctions à q variables d'origines, ou autrement dit les projections sont simplement des sélections de q variables d'origine. Plus généralement, on a des combinaisons linéaires des variables d'origines, ce qui est important c'est que les f_k sont des fonctions à q variables.

Comment cela permet de résoudre le problème de la dimension? On passe en fait d'1 problème de dimension d à K problèmes de dimension q. Si on impose un peu de régularité (type lipschitz) sur chaque f_k, alors

$$\|f_k - f_{\theta_k}\|_\infty \leq C_k \sqrt{q} n^{-1/q} \tag{24}$$

et donc par addition sur k on obtient finalement

$$\|f - f_\theta\|_\infty \leq CK\sqrt{q} n^{-1/q} \tag{25}$$

($C = \max C_k$). C'est-à-dire que l'on a augmenter la vitesse de convergence de $n^{-1/d}$ à $n^{-1/q}$. **Ainsi, si le problème initial est séparable en petits problèmes de faible dimension on a quelque chose de gérable.**

Cette façon de procéder de séparation d'un problème en sous problèmes a une application fondamentale dans l'estimation d'une densité de probabilité, soit $f(x) = \log p(x)$. L'hypothèse de séparabilité revient à faire l'hypothèse que

$$p(x_1, \ldots, x_d) = \Pi_{k=1}^{K}\ p_k(x_{i_1^k}, \ldots, x_{i_q^k}) \tag{26}$$

c'est-à-dire que l'on traite des modèles de Markov (ou modèles graphiques). Dans ce contexte, on peut séparer les groupes de q variables qui incluent par exemple x_1, et ceux qui ne l'incluent pas. Pour ceux qui incluent x_1, ils décrivent des processus en interaction avec x_1; les autres décrivent des processus indépendants de x_1. Donc, pour comprendre l'évolution en fonction de x_1, on se concentre sur les premiers processus. C'est donc une approche "locale".

Dans le cas des images, la séparation du problème peut se traduire par le découpage de l'image originale en petits patchs qui seront considérés comme indépendants les uns des autres, à savoir que seuls les pixels dans chaque patch sont en interaction les uns avec les autres. Ainsi, on peut appréhender la classification de l'image (cf. chien vs chat) comme la somme des vraisemblances sur les petites images car on pense avoir suffisamment de puissance de reconnaissance de features sur les patchs. Si on peut faire cela alors le problème de dimension $10^3 \times 10^3$ passe par la résolutions de problèmes de dimension 8×8. Ce type d'approche était largement majoritaire avant les années 2010, où l'on utilisait des descripteurs locaux invariants pour réduire au-delà de 64 la dimensionalité du problème par la méthode SIFT (*scale-invariant feature transform* datant de 1999), sur laquelle était appliqué des méthodes de classifications. C'était l'état de l'art à cette époque.

Pour le cas de reconnaissance de musiques, de phonèmes etc, on peut diviser le temps en petits intervalles typiques de 25 ms (jusqu'en dans les années 2010). On a par exemple 200 d'échantillons, si on échantillonne à 8 kHz, et on calcule la décision finale comme une somme les décisions individuelles. On peut faire des choses plus subtiles (processus de Markov) mais pour la classification de type de musique, on faisait la somme d'évidence locale. Les descripteurs locaux étaient des MFCC (*Mel-Frequency Cepstral Coefficients*) qui étaient conçus sur la construction d'invariants qui sont stables par déformations (Voir Sec. sec-2020-mfc1). En parallèle, il y avait les réseaux de neurones mais clairement dans les années avant 2010, ils ne marchaient pas bien pour ces problèmes.

On peut également se poser la question de cette technique de séparabilité en chimie quantique. Les x sont les positions et charges et la fonction à trouver est l'énergie quantique. Les liaisons de covalences sont celles qui relient les atomes entres-eux par partage d'électrons. On peut raisonnablement penser que l'énergie est dominée par des termes qui décrivent la relation locale de voisinage entre atomes. On sait que cela n'est pas exacte, car il y a des termes à plus longue distance mais on peut les incorporer. Cependant, le gros point noir, ce sont les termes quantiques, mais on peut utiliser des descripteurs qui renseignent sur la structure locale (comme en linguistique les *bags of words*). Ça peut marcher mais pas toujours.

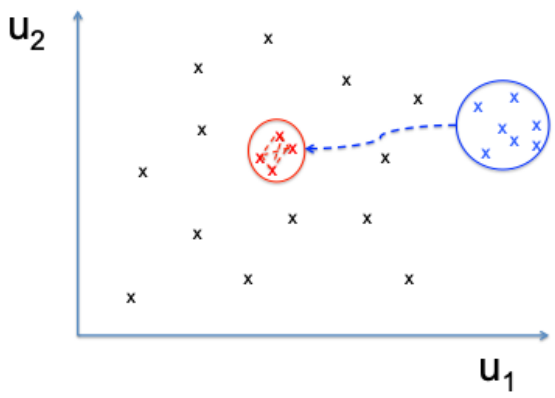

FIGURE 12 – Illustration des interactions à courte portée au sein du groupe rouge, et à longue portée entre le groupe rouge et le groupe bleu.

4.3.2 Séparabilité des échelles

En fait, l'article d'H. Simon donne une piste, car il y a des interactions globales qui ne peuvent pas être négligées. Imaginons des interactions sur un réseau social, on peut se dire que typiquement ayant une dizaine de personnes avec lesquelles on est en relation, notre sphère d'influence est circonscrite à ces 10 personnes et peu importe des événements qui se passent à l'autre bout de la planète. Or, c'est bien un fait que la géopolitique aux antipodes a un impact sur nos propres décisions. Donc, des interactions à longue distance entre groupes d'individus sont à prendre en compte. Ce raisonnement, on peut le tenir aussi pour la classification des chiens et des chats, car la mettre en œuvre uniquement à partir de patchs 8x8 n'est sans doute pas si simple que cela. De même, pour reconnaitre le compositeur d'une œuvre musicale on ne peut se contenter de faire des analyses sur 25 ms, il faut faire le lien entre les différents échantillons.

Envisageons donc une structuration hiérarchique: pourquoi cela apporte une solution? En fait, on va à la fois se concentrer sur l'étude des interactions entre tous les agents/entités à petites échelles, et ne considérer que les interactions à grandes échelles entre les groupes qui sont pris comme des formes globales interagissantes. Une illustration de ce propos est donné dans la figure 12. Ce type de schéma peut se concevoir par exemple en chimie/physique, voire en sociologie.

Si on peut traiter le problème initial de la sorte alors on n'a pas à traiter d variables

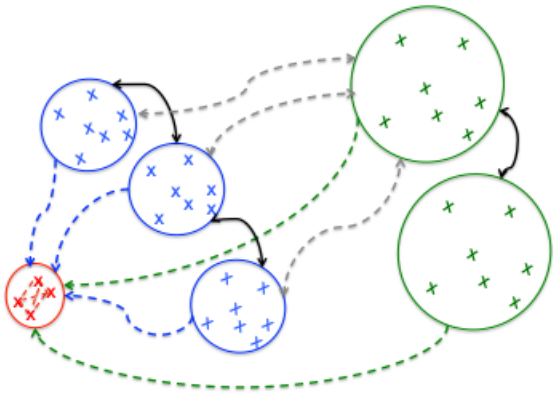

FIGURE 13 – Illustration du rôle des interactions entre groupes à différentes échelles.

mais typiquement $\log d$ groupes. Or, passer de d à $\log d$ résout le problème de la malédiction de la dimensionalité. **Mais il y a une hypothèse qui ne va pas marcher: c'est la forme de la décomposition (Eq. 22) en somme de problèmes indépendants.** Or, on peut concevoir qu'il faille considérer pour certains problèmes toutes les interactions entre tous les groupes pour comprendre ce qui se passe pour un groupe donné (voir figure 13). C'est la grosse difficulté car même si les groupes ont des interactions faibles entres eux, 1) la somme de leurs interactions n'est pas forcément négligeables vis-à-vis des interactions à courtes distances et 2) elles dominent selon H. Simon la dynamique lente du système global.

Ceci dit, on va mettre en place la **séparation d'échelle** avec un outil mathématique qui est **l'analyse en ondelettes** et on se posera la question de savoir comment capturer les interactions d'échelles.

4.3.3 Notions générales sur les groupes

Cette notion dit du système qu'il y a des états équivalents que l'on connait à l'avance. Pour connaitre la régularité de $f(x)$ on peut étudier sa régularité par rapport à des transformations [31]. Dans un premier temps, on peut utiliser des transformations *locales* pour savoir si la fonction est continue, dérivable, etc., mais ce qui nous intéresse ici ce

31. NDJE: voir Cours 2019 Sec. 3.5

sont des classes de transformations globales, c'est-à-dire des groupes de symétries de f

$$G = \{g \ / \ \forall x \in \Omega, \ f(g.x) = f(x)\} \tag{27}$$

Les fonctions g préservent les lignes de niveaux. En effet, soit l'ensemble Ω_t tel que

$$\Omega_t = \{x \ / \ f(x) = t\} \tag{28}$$

c'est bien une ligne de niveau, et Ω_t est invariant par g.

Def. d'un groupe: Soit une opération de $G \times G \to G$, telle que $(g_1, g_2) \to g_1.g_2$, elle définit la structure (morphisme) de groupe si on a les propriétés suivantes:

— associativité: $\forall g_1, g_2, g_3 \in G$, on a $(g_1, g_2).g_3 = g_1.(g_2.g_3)$

— élément neutre : $\exists Id \in G$, tel que $g.Id = Id.g = g$

— l'inverse : $\forall g \in G$, $\exists g^{-1} \in G$ tel que $g.g^{-1} = g^{-1}.g = Id$

— commutativité (option; groupe abélien): $\forall g_1, g_2 \in G$, $g_1.g_2 = g_2.g_1$

L'idée géniale de Gallois est d'avoir su identifier que pour étudier des fonctions, on pouvait le faire à travers leurs groupes de symétrie. Par exemple, dans le cas de la recherche de solutions d'une équation polynomiale, $P(x) = 0$, c'est la ligne de niveau 0, on peut regarder les transformations qui transforment une solution en une autre solution. Pourquoi, les solutions d'une courbe de niveau forment-elles une structure de groupe? Soit donc $x \in \Omega_t$ et $g \in G$ alors $g(x) \in \Omega_t$. On vérifie alors aisément l'associativité. Il est clair aussi que l'identité est un élément de G et quelle commute avec tous les éléments $g \in G$. Et si on se restreint aux opérateurs inversibles donc si g transforme une solution x_0 en une autre solution x_1, son inverse opère la transformation de x_1 en x_0, et on a bien alors une structure de groupe. Ce type d'approche est centrale dans l'étude des équations différentielles (partielles): si on a une solution, quelles sont les opérations qui préserve la solution?

Exemple la translation sur une grille discrète $u \in \mathbb{Z}^2$, $g.x(u) = x(u - g)$. Une fois que l'on a un groupe, on se pose la question de ses générateurs et la dimension du groupe est le nombre de ses générateurs: $\{g_k\}_{k \leq P}$. Dans le cas d'un groupe commutatif, un élément du groupe peut se décomposer selon $g = g_1^{n_1}.g_2^{n_2} \dots g_k^{n_k} \dots g_P^{n_P}$. Mais plus généralement (sans commutativité), l'action des générateurs doit se faire séquentiellement en respectant un ordre.

Avant de passer au groupe continu, demandons-nous à quoi sert cette notion de groupe? Cela opère une **réduction de la dimensionalité**. Si on dispose de l'information *a priori* suivante: on connait un sous-groupe H du groupe de symétrie G que peut-on en conclure? En passant, si on connait la totalité du groupe de symétrie de la fonction f, on connait *de facto* toutes les lignes de niveaux de la fonction, donc on connait toutes les solutions aux équations de type $f(x) = t$, donc on connait la topologie de la fonction f, en bref on la connait totalement. Si part contre, on ne connait que $H \subset G$, ce qui est le cas concret en pratique: on ne peut connaître tout G, mais dans le problème de classification d'images, il est clair que la translation, la rotation, le flip d'un objet ne change pas son étiquette. Ainsi, pour $g \in H$, x et $g.x$ appartiennent à la même classe d'équivalence, et **on va utiliser le quotient de Ω par H, noté Ω/H**. Si on prend $x_0 \in \Omega/H$, il définit une **classe d'équivalence** H_{x_0} tel que

$$H_{x_0} = \{x \in \Omega \ / \ g \in H \text{ tq } g.x = x_0\} \tag{29}$$

et on ne peut distinguer en quelque sorte $f(x \in H_{x_0})$ de $f(x_0)$. Si x_0 est une image, $f(x_0)$ son label (cf. chien/chat), alors si on la translate on obtient $x = g(x_0) \in H_{x_0}$ et $f(x) = f(x_0)$ le label est le même. **Comme le label est le même pour tous les éléments de H_{x_0}, on peut réduire le nombre de variables pour ne pas prendre en compte la variabilité au sein de la classe d'équivalence**. C'est bien de la réduction de dimensionalité.

De combien de variables va-t'on réduire en opérant la quotientisation: typiquement dans un cadre continue on s'attend à $\dim(\Omega/H) = \dim(\Omega) - \dim(H)$. Ça veut dire que pour être utile **la dimension de H doit être grande**. Par exemple, si on prend une translation à pas fixe dans une image, sa dimension est de 2, mais face aux 10^6 variables d'une image 1024×1024, la réduction de dimension est totalement anecdotique. Donc, **l'invariance par translation à elle seule n'est pas une information suffisamment forte**.

4.3.4 Difféomorphisme, groupe des déformations

Introduisons les idées des groupes de Lie avec comme fil conducteur la translation continue. On est donc en présence de translation g dans $\mathbb{R}^2 = G$. G est une variété différentiable de dimension P. Les propriétés de groupe sont les mêmes que celles des groupes "classiques", mais il y a maintenant une notion nouvelle qui est celle de **transport**

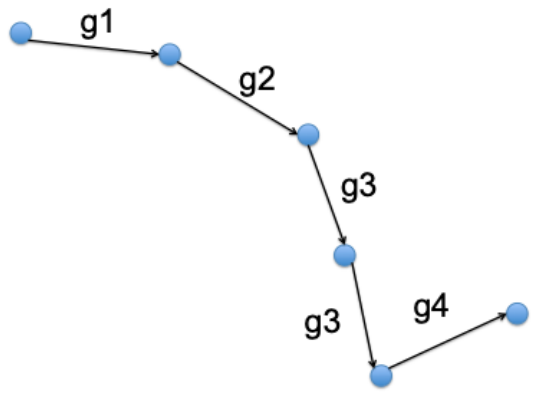

FIGURE 14 – Illustration de la notion de transport.

(Fig. 14). Un élément x_0 par action successive des générateurs (ex. ceux du groupe de symétrie) se transporte en x_1, x_2, etc et à chaque fois $f(x_i) = f(x_0)$ est inchangé, de plus on peut le transporter sur possiblement de grandes distances. La grande différence est que le transport est continue (voire différentiable), et en prenant toutes les façons de transporter x_0 avec les générateurs (du groupe des symétries) on obtient une **surface différentiable d'iso-label que l'on appelle *l'orbite* de** x définie par $O_x = \{g.x\}_{g \in G}$.

Si on veut caractériser la surface différentielle O_x, on étudie **les hyper-plans tangents** qui donnent la dimension de la variété, et qui sont définis par des générateurs infinitésimaux. L'Algèbre de Lie est l'algèbre qui fait passer d'une transformation à une autre infiniment proche. Dans le cas d'une image à 2 dimensions $u = (u_1, u_2)$, alors pour une petite transformation g appliquée à $x(u)$ on obtient:

$$g(x(u)) = x(u - g) = x(u) - \nabla x(u).g + \cdots = x(u) - \left\{ \frac{\partial x}{\partial u_1} g_1 + \frac{\partial x}{\partial u_2} g_2 \right\} + \ldots \qquad (30)$$

Donc, localement on un groupe de dimension 2, et les générateurs de l'algèbre de Lie sont les dérivées partielles $(\partial x/\partial u_1, \partial x/\partial u_2)$ et la direction de translation dans le plan tangent est donnée par les coordonnées (g_1, g_2).

Donc en continu, le groupe de Lie est une variété, dont les plans tangents sont générées par des générateurs de l'algèbre de Lie de dimension fixée. **La particularité des variétés générées par un groupe de Lie, est que les plans tangents sont tous identiques car**

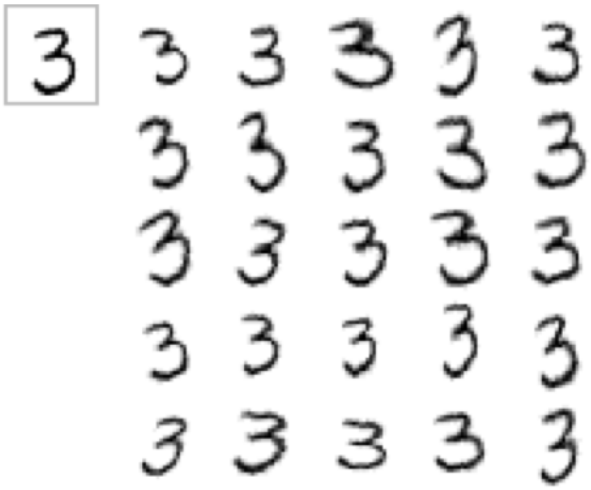

FIGURE 15 – Déformation d'une image d'un 3.

générés par les mêmes générateurs. Ainsi, les groupes de Lie sont en quelque sorte des structures extrêmement rigide. Quelques exemples de groupe de Lie (dimension finie):

— une translation: $g.x(u) = x(u - g)$ avec $g \in \mathbb{R}^2$ (dim: 2)

— une rotation: $g.x(u) = x(r_g.u)$ avec $g \in [0, 2\pi]$ (dim: 1)

Le groupe des déformations (figure 15) est intéressant car agissant sur le chiffre 3 par exemple, on peut penser que le résultat de la classification ne devrait pas changer à la suite de petites déformations. Cette information *a priori* est donc intéressante à pouvoir être traduite dans la classe de fonctions f_θ qui classent les digits. Or, une déformation agissant sur chaque pixel, est potentiellement régit par beaucoup de paramètres, ce qui induit donc **une dimensionalité du groupe des déformations très importante**: c'est ce que l'on recherche. Cela va structurer profondément le problème de classification, mais également d'autres types de problèmes en traitement d'images, de son, en physique/chimie...

Le groupe est celui des difféomorphismes et un élément g du groupe agit selon la relation suivante:

$$g.(x(u)) = x(g(u)) \qquad u \in [0, 1]^2 \tag{31}$$

En fait g agit sur les variables sous-jacentes de x, à savoir u qui est une quantité de basse dimension. On va imposer que g soit une fonction continument dérivable C^1: $g : [0, 1]^2 \to [0, 1]^2$. Comme en chaque point (pixel) d'un 3, on a une infinité de possibilités

de transformation de ce pixel, la dimensionalité du groupe est infinie.

5. Séance du 12 Février

5.1 Rappels introductifs

Cette séance est la continuité de la précédente. Après un petit rappel, S. Mallat reprend le fil sur le thème des symétries de $f(x)$.

Les trois notions, séparabilité, symétrie et "sparsité" ("les trois S"), sont toutes liées à la régularité de la fonction à apprendre $f(x)$, avec x évoluant en grande dimension. Nous avons vu que ces trois notions permettent une réduction de dimension, ce qui permet de combattre la malédiction de la dimensionalité.

La séparabilité permet de découper le problème initial en sous problèmes qui traitent des "blocs" de x qui ont une interaction locale et peu d'interaction entre blocs. A travers la notion de symétrie, on regarde les invariants de la fonction $f(g.x) = f(x)$ et les lignes de niveau sont invariantes par g qui ont une structure de groupe H [32]. Cette structure de groupe permet de quotienter Ω par H et ne garder que les effets de la fonction f sur les classes d'équivalence, c'est-à-dire que l'on réduit la dimension du problème selon $\dim(\Omega/H) = \dim(\Omega) - \dim(H)$ (sauf cas pathologique). Et plus nous pouvons obtenir d'information *a priori* à l'aide d'un groupe de très grande dimension, plus la réduction du problème initial est efficace.

Quelques exemples de groupe:

— Dans certains problèmes, comme l'analyse d'images médicales, il est judicieux de renormaliser les valeurs des pixels [33] dans des intervalles typiques $[0, 255]$ ou $[0, 1]$: la transformation est du type $g.x(u) = \theta\ x(u)$ c'est le groupe multiplicatif avec $\theta \in \mathbb{R}$;

32. Rappel: si on connait toutes les symétries de f, c'est-à-dire le groupe G, alors on connait la fonction; ce qui n'est pas le cas donc on note H le groupe des symétries connues *a priori*.

33. NDJE: il est parfois pas possible de faire cette transformation, car le problème doit prendre en compte les différences de valeurs d'une image à l'autre: ex. pour mesurer la distance d'une galaxie en utilisant des images CCD, il faut typiquement non seulement tenir compte de l'extension de la galaxie qui pourrait se faire à valeurs de pixel renormalisées, mais aussi de l'intensité des pixels qui renseigne sur le flux reçu.

— une translation: $g.x(u) = x(u - \theta)$ avec dans une image $\theta \in \mathbb{R}^2$; idem pour la rotation, etc
— une déformation: $g.x(u) = x(\theta(u))$, c'est une fonction représentant une action locale que l'on suppose inversible et C^1, $\theta : \mathbb{R}^2 \to \mathbb{R}^2$. C'est le groupe des difféomorphismes dont la dimension est infinie.

Les transformation de x par les actions des générateurs (figure 14) définissent une hypersurface appelée l'orbite de x, notée O_x qui a le même "label" $f(x)$. Pour décrire l'orbite, on étudie ses hyper-plans tangents, et la propriété remarquable est qu'ils sont tous identiques, ce qui fait de la structure de groupe une structure très rigide.

5.2 La représentation $\Phi(x)$

On va utiliser l'information *a priori* sur le groupe de symétries de $f(x)$ pour définir la représentation $\Phi(x)$ qui est l'étape avant la classification/régression finale (figure 2):

$$\tilde{f}(x) = \langle \Phi(x), w \rangle = \sum_k w_k \phi_k \tag{32}$$

Or, on veut que $\tilde{f}$ soit une bonne approximation de f la fonction cherchée, à ce titre on va imposer que $\tilde{f}$ ait les mêmes invariants $g \in G$ que f, ce qui va se faire en imposant que G soit un groupe de symétries de Φ. A ce stade deux cas de figures se présentent.

Soit le groupe G est de basse dimension et il est connu (ex. translation, rotation, multiplication), alors on va construire $\Phi(x)$ en conséquence. Mais gardons à l'esprit que ce type de groupes ne change pas le problème initial de travailler en grande dimension.

Soit le groupe G est très grand, voir de dimension infinie: dans ce cas le problème est que l'on ne connait pas le groupe. Reprenons le cas des déformations de chiffres, par exemple la figure 15 pour le 3 mais il faut imaginer le même type de déformation pour les autres chiffres; or, si on n'y prend garde une grande déformation peut faire passer un 3 en 5, un 1 en 2, etc. Donc, si on note $\bar{G}$ le groupe "total" des difféomorphismes, il va ne falloir considérer qu'un sous-groupe $G \subset \bar{G}$. Mais on ne peut le connaitre totalement: certes les petites déformations appartiennent sans nul doute à G, mais à partir de quelle "ampleur" une déformation n'en fait pas partie, là est la question. On a une **information partielle *a priori***: le groupe de symétrie appartient à $\bar{G}$. C'est ce qui motive l'apprentissage et l'on

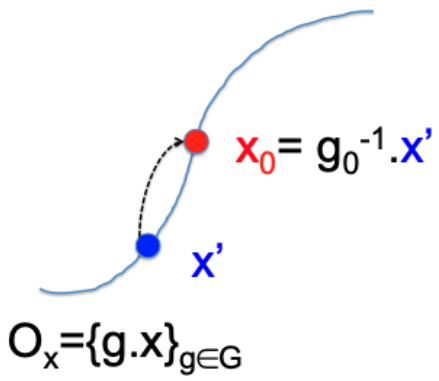

FIGURE 16 – Transformation de tout $x' \in O_x$ en un point x_0 de l'orbite invariant par une nouvelle action.

va "apprendre" G à travers w, le paramètre de la classification. Ainsi, on n'impose pas à $\Phi(x)$ d'être invariant par $\bar{G}$, car il contient des déformations beaucoup trop grandes qui induiraient des erreurs énormes. **Par contre, on peut imposer des formes faibles qui sont des linéarisations; et les déformations qu'il faut supprimer, on les apprend à travers l'apprentissage de** w**.**

5.3 L'échec des invariants canoniques

La manière traditionnelle utilisée pour introduire les invariants n'est pas la bonne dans notre cas (sic). En effet, traditionnellement (ou la plus simple) on utilise les **invariants canoniques**. **L'idée est de "renormaliser" les points** $x' \in O_x$ **de l'orbite de** x **en un point** x_0 **invariant par nouvelle transformation** (figure 16). Considérons le groupe multiplicatif qui sert à normaliser les pixels d'une image, ainsi tout $x' \in O_x$ subit la transformation:

$$x_0 = g_\theta.x' = \theta \times x' \quad \text{tq} : x'(u) \rightarrow \frac{x'(u)}{\sum_u |x'(u)|} = x_0(u) \tag{33}$$

donc

$$\theta_0 \equiv \sum_u |x'(u)| \Rightarrow x'(u) \rightarrow \theta_0^{-1} \times x'(u) = x_0(u) \tag{34}$$

Après renormalisation $\sum_u |x_0(u)| = 1$, donc tout facteur multiplicatif ne change plus la configuration des x. Dans le cas de la translation,

$$g_\theta.x(u) = x(u - \theta) \tag{35}$$

et par exemple on veut "recaler" toutes les images par rapport au centroïde des pixels:

$$\theta_0 \equiv \frac{\int u|x(u)|du}{\int |x(u)|du} \tag{36}$$

on définit alors $x_0(u) = g_{\theta_0}^{-1}.x(u) = x(u + \theta_0)$. Le nouveau centroide est égal à 0; on a bien "renormalisé" les éléments de l'orbite de x. Attention, il faut bien voir que tous les points de l'orbite de x sont renormalisés au même point x_0, mais que des x d'orbites différentes sont renormalisés en des points différents: imaginons un "3" dans une image, s'il est translaté (tourné, ...) on veut le recalé de tel façon que le centroïde soit au centre de l'image, une fois trouvé le "3" de référence, tous les nouveaux "3" sont renormalisés pour se conformer au "3" de référence; mais pour les "2" la transformation est différente (cf. l'orbite des "3" est différente de l'orbite des "2").

Donc l'idée de l'invariant canonique, c'est d'estimer le paramètre de la transformation θ_0. **Quand on a un grand groupe, peut-on faire de même?** Prenons le cas des difféomorphismes, quelle serait l'idée (voir les travaux de Grenender et Miller des années 1997: *Deformable templates*)? On décrit tous les objets à partir des déformations d'un objet de référence. Exemple: on prend les chiffres de fontes normalisées, et on décrit un chiffre manuscrit comme la transformation de son archétype. Pour identifier un chiffre manuscrit, on choisit l'archétype qui est identifié par la plus petite déformation qui fait passer de l'un à l'autre. On essaye d'estimer la déformation par rapport à un template de référence. Cela marche bien en imagerie médicale: si on veut décrire l'atlas du cerveau, on peut utiliser les images IRM effectués sur de nombreux sujets et on "recale" les images qui donnent un atlas de référence, puis de nouvelles images d'un patient sont identifiées par rapport à cet atlas. **Le point crucial est que l'on peut définir cet atlas de référence.** Dans le cas d'images plus hétéroclites, il va être difficiles voir impossible d'obtenir cet atlas, soit parce qu'il n'y a pas d'objets de référence, soit parce qu'il devrait y en avoir un nombre très grand; de plus il faudrait par la suite pouvoir estimer toutes les déformations de l'objet à identifier par rapport à tous les templates de référence! **Le programme ne peut être mené**

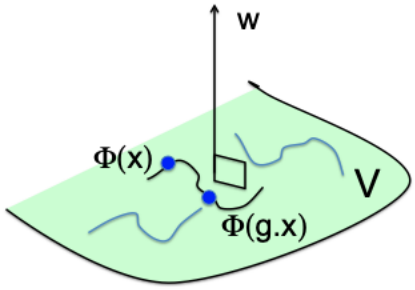

FIGURE 17 – Le transformé de x via la représentation Φ, ainsi que le transformé de $g.x$ sont sur le même plan $V \perp w$.

à termes, cette façon de procéder est un échec (sauf cas particulier). En définitive, on n'est pas demandeur de calculer la déformation, ce que l'on cherche c'est de savoir si on a affaire à un 1, 2, 3, etc dans le cas de la reconnaissance des digits. Donc, **au lieu de calculer la déformation, on va l'éliminer**.

Quelles sont les conditions pour que cela marche? Si on reprend la représentation $\Phi(x)$, on aimerait avoir:

$$\tilde{f}(x) = \tilde{f}(g.x) \Rightarrow \langle \Phi(x), w \rangle = \langle \Phi(g.x), w \rangle \Rightarrow \langle \Phi(x) - \Phi(g.x), w \rangle = 0 \tag{37}$$

donc

$$\Phi(x) - \Phi(g.x) \in V \perp w \tag{38}$$

comme schématisé sur la figure 17. Si donc V est un hyper-plan, cela implique alors de **linéariser les transformations**, donc typiquement on prendra des g infinitésimaux.

5.4 Linéarisation de l'action de groupe

On va se concentrer sur le groupe des difféomorphismes (déformation), car il joue un rôle très important que cela soit pour les images dont les déformations reflètent toute une variété de situations de vue d'un objet par exemple, mais également pour les voix dont la variabilité reflète celle des timbres, des rythmes etc, de l'élocution d'un sujet, etc. Que signifie linéariser une petite déformation? On considère des actions de groupe avec une

petite transformation τ:

$$g.x = x + \tau.x \tag{39}$$

Notons que localement l'hyperplan tangent à l'orbite de x est donné par les vecteurs τ indépendants (algèbre de Lie). On réécrit la transformation précédente selon

$$g.x(u) = x(u - \tau(u)) = x(u) - \nabla x(u).\tau(u) \tag{40}$$

Les $\tau(u)$ représentent **le champ de déplacements**: se déplacer le long d'un difféomorphisme, localement est équivalent à faire un déplacement dans le plan tangent. Donc, les générateurs de l'hyper-plan tangent est un champ de déplacements. Si $x(u)$ est régulier (C^1) alors par Taylor on obtient la seconde égalité ci-dessus (nb. $u = (u_1, u_2)$ dans une image). Cependant, pour mieux comprendre comment on peut obtenir de manière générale des invariants, on va un peu plus détailler cet expression (Eq. 40). On va se permettre d'écrire l'action de τ comme une action "globale" et une petite action "locale", selon:

$$\tau(u) \approx \tau(u_0) + \nabla\tau(u_0)(u - u_0) \tag{41}$$

Notons que l'on a

$$\nabla\tau(u) = \begin{pmatrix} \partial_{u_1}\tau_1(u) & \partial_{u_2}\tau_1(u) \\ \partial_{u_1}\tau_2(u) & \partial_{u_2}\tau_2(u) \end{pmatrix}$$

Donc, par développement simple on obtient:

$$\begin{aligned} x(u - \tau(u)) &= x\left(u - \tau(u_0) - \nabla\tau(u_0)\ (u - u_0)\right) \\ &= x\left(\underbrace{\left(\mathbb{I} - \nabla\tau(u_0)\right)(u - u_0)}_{\text{déformation locale}} + \underbrace{u_0 - \tau(u_0)}_{\text{translation globale}}\right) \end{aligned} \tag{42}$$

avec les deux types d'actions de la déformation $\tau(u)$. De plus, on peut définir une taille du groupe des petits difféomorphismes selon:

$$|g|_G = \underbrace{\|\tau\|_\infty}_{\text{translation}} + \underbrace{\|\nabla\tau\|_\infty}_{\text{déformation}} \tag{43}$$

avec $||\tau||_\infty = \sup_u |\tau(u)|$ et $||\nabla\tau||_\infty = \sup_u ||\nabla\tau(u)|| < 1$, c'est-à-dire que la plus grande valeur propre du jacobien est plus petite que 1 (inversibilité). Ce que cela dit, c'est que la taille de l'action de groupe des déformations est égale à la somme de la valeur maximale de la translation globale et de la valeur maximale de la déformation locale. Si par exemple $\tau(u) = \varepsilon u$, soit une dilatation infinitésimale, alors $||\nabla\tau||_\infty = \varepsilon$, mais donc en principe on dispose d'une formulation générale qui comprend tous les types de déformations.

Peut-on et comment linéariser les déformations via Φ? Finalement, on essaye de construire des conditions sur Φ et nous verrons que cela nous mène aux réseaux de neurones et à la problématique de la séparation d'échelles. En reprenant l'équation 37 et en injectant le développement (Eq. 40),

$$\Phi(x) - \Phi(g.x) \approx \Phi(x) - \Phi(x) - \nabla\Phi(x)(\tau.x) = -\nabla\Phi(x)(\tau.x) \tag{44}$$

Donc, au premier ordre l'ensemble V se traduit par [34]:

$$V = \{\nabla\Phi(x)(\tau.x) \ / \ \forall x \in \Omega, \forall \tau \text{ géné. G}\} \tag{45}$$

Quelles sont les inconnues? 1) le groupe "exact" G même si on ne le rappelle pas, 2) les générateurs ce qui revient à 1), et donc on n'a pas vraiment de connaissance sur V. **En revanche, avec l'apprentissage on va déterminer w (tout du moins celui accessible via l'apprentissage, lapalissade), donc indirectement on identifie V, donc les générateurs et donc un sous-groupe H des symétries de f.** Cependant, pour que cela marche, il faut pouvoir calculer la "formule de Taylor" sur $\Phi(x)$, donc Φ **doit être différentiable relativement à l'action du groupe $\tau.x$.** C'est là où est le problème! Tout du moins il faudrait s'assurer que $\Phi(x)$ soit Lipschitz, c'est-à-dire que sous l'action d'une (petite) transformation g du groupe G [35]:

$$||\Phi(x) - \Phi(g.x)|| \leq C \underbrace{d(g, \mathbb{I})}_{|g|_G} ||\Phi(x)|| \tag{46}$$

34. ici on fait apparaître G le groupe de symétries "exacts" de f alors que l'on a accès qu'à un sous-groupe

35. NDJE: Pour comprendre cette formulation de la condition de Lipschitz, prenons $\Phi(x) = \langle x \rangle$, c'est-à-dire la moyenne de x, et prenons l'action d'une dilatation infinitésimale $g = \mathbb{I} + \varepsilon$, alors $||\Phi(x) - \Phi(g.x)|| = ||\langle x \rangle - \langle (1+\varepsilon)x \rangle = \varepsilon||\langle x \rangle|| = d(g, \mathbb{I})||\Phi(x)||$.

Donc, cela dit que la différence entre l'action de Φ sur le transformé de x (cf. $g.x$) doit être de la taille de g. Si le problème est invariant par translation (cas simple), $|g|_G = \|\nabla\tau\|_\infty$, donc il faut que la variation de Φ soit petite devant le gradient de la déformation (ici translation locale). Plus généralement, en utilisant Eq. 43 on peut écrire pour de petites déformations la condition de Lipschitz pour Φ selon:

$$||\Phi(x) - \Phi(g.x)|| \leq C||\Phi(x)|| \left(\underbrace{\|\tau\|_\infty}_{\text{translation}} + \underbrace{\|\nabla\tau\|_\infty}_{\text{déformation}} \right) \tag{47}$$

Récapitulons à ce stade: on ne connait pas le groupe G donc ses invariants, donc on linéarise pour éliminer les petites transformations, ce qui n'est possible que si Φ est Lipschitz, alors on pourra identifier les générateurs par apprentissage et "quotienter" l'espace Ω.

Maintenant, revenons aux déformations après élimination des translations, il faut que $|g|_G \ll \|\nabla\tau\|_\infty$. C'est une condition que l'on va systématiser pour toutes les déformations car sinon on ne pourra pas les identifier pour les quotienter. **Or, la propriété que Φ soit Lipschitz (Eq. 46) est difficile à obtenir.** En effet, une déformation locale, c'est une dilatation, donc il va falloir baser la représentation sur des dilatations, c'est-à-dire faire de la **séparation d'échelles que l'on obtient par transformée en ondelettes.** Et on va se retrouver typiquement sur des structures qui ressemblent aux réseaux de neurones.

5.5 Étude d'un recalage

Soit le centroïde de x que l'on note θ_0, un recalage produit alors:

$$g.x(u) = x(u - \theta_0) \tag{48}$$

Le problème est que $x(u)$ n'est pas régulier: ex. dans une image, il y a des discontinuités naturelles qui sont dues aux différents objets/paysage/textures présents dans la scène; idem dans un échantillon de sons, etc... Si l'on fait suivre une petite dilatation après un recalage, la distance entres deux bumps passe de d à $d + 2d\varepsilon$ (figure 18). La conséquence est que les bumps avant et après dilatation n'ont pas de recouvrement spatial. Alors, la

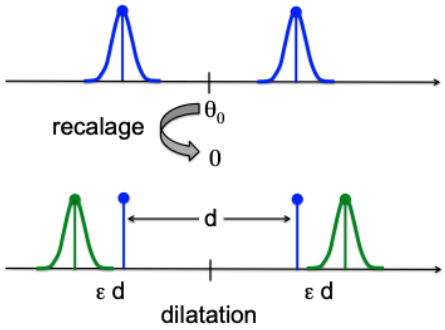

FIGURE 18 – Effets d'une étape de recalage "global" suivie d'une petite "dilatation" sur 2 bumps distants d'une distance d.

distance entre $\Phi(x)$ et $\Phi(g.x)$ est égale à :

$$||\Phi(x) - \Phi(g.x)||^2 = ||\Phi(x)||^2 + ||\Phi(g.x)||^2 - 2\int \Phi(x)\Phi(g.x)dx \approx 2||\Phi(x)||^2 \tag{49}$$

Et donc la différence $||\Phi(x) - \Phi(g.x)||$ peut être très grande en définitive. **En fait, les irrégularités (hautes fréquences) sont très instables par dilatation.** C'est pour cette raison qu'il n'est pas suffisant de faire des recalages pour faire de la reconnaissance de forme (ex. visage). Nous retrouverons ce phénomène dans le domaine de Fourier à la section 6.3.2.7.

5.6 Autre invariant, la covariance de groupe

Si on dispose de l'orbite de x (O_x), essayons d'imposer un **invariant linéaire**, c'est-à-dire envisageons les combinaisons du type $\sum_{g\in G} \alpha_g g.x$. Or, cette combinaison est invariant *ssi* les α_g sont identiques, donc cela revient à faire une **moyenne**:

$$\sum_{g\in G} g.x \tag{50}$$

Pour une translation cela revient à faire la somme sur toutes les translations possibles

$$\sum_{\theta} x(u-\theta) \tag{51}$$

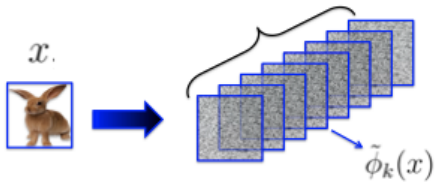

FIGURE 19 – Étape dans un réseau de neurones qui fait passer l'image d'entrée x à un ensemble d'image $\left\{\tilde{\phi}_k(x)\right\}_k$ (nb. pas de même dimension en général).

qui n'est autre que la moyenne du signal x qui est bien un invariant par translation [36]. De même, la moyenne du signal est invariante par dilatation [37]. **Donc, certes on a un invariant par translation et dilatation, à savoir la moyenne du signal, mais on a perdu toute information sur la structure de celui-ci.** De ce point de vue, l'invariant canonique qui est un représentant de l'orbite de x, garde toute la structuration de l'orbite. Tandis que la moyenne, n'est pas spécifique à une orbite particulière, cf. une infinité de signaux ont la même moyenne. Donc, **un invariant linéaire est beaucoup trop simple**, voire naïf. L'idée va être de moyenner beaucoup de canaux issus de x.

Soit $\tilde{\phi}(x)$ définit par l'ensemble des images à une étape d'un réseau de neurones (figure 19): chaque $\widetilde{\phi_k}(x)$ ($k \in \{1, \ldots, K\}$) est une image indexée par u. Donc maintenant, on dispose d'une collection d'images et on va essayer de les rendre invariantes selon la moyenne spatiale:

$$\sum_{\theta}[\tilde{\phi}(x)](u-\theta) = \sum_{g\in G} g.[\tilde{\phi}(x)] \equiv \Phi(x) \tag{52}$$

C'est ce que l'on appelle un **pooling**, et c'est typiquement ce qui est effectué à la fin du réseau, juste avant la partie purement *fully-connnected* du classificateur, on fait un *flat* car on a déjà tout moyenné sur l'espace.

La question est de savoir si la transformation 52 est invariante, c'est-à-dire que si $g_0 \in G$ alors $\Phi(g_0.x) = \Phi(x)$? La condition est que $\Phi(x)$ soit **équivariante par l'action du**

36. NDJE: pour s'en convaincre prenons un signal $x(t)$ périodique sur $[0, 2\pi]$, simplement la somme sur θ se traduit par une intégrale $(2\pi)^{-1}\int_0^{2\pi} x(u-\theta)d\theta$ qui par changement de variable $u' = u - \theta$ et périodicité du signal, donne $(2\pi)^{-1}\int_0^{2\pi} x(t)dt$ qui n'est autre que la moyenne du signal.

37. NDJE: avec le signal de la footnote 36, on calcule la moyenne sur θ de $x(u\theta)$ sur une période $[0, 2\pi/u]$ ce qui redonne la moyenne du signal.

groupe G (nb. vrai pour tous les $\widetilde{\phi_k}(x)$), c'est-à-dire que

$$\boxed{\Phi(g.x) = g.\Phi(x)} \tag{53}$$

en gros Φ et g commutent dans leur action sur x. Notons bien la différence entre **invariance** et **équivariance** [38]

$$\begin{cases} f(g.x) = f(x) & \text{invariance} \\ f(g.x) = g.f(x) & \text{équivariant} \end{cases} \tag{54}$$

Par exemple, si on translate x alors les $\widetilde{\phi_k}(x)$ doivent être les mêmes que l'image d'origine mais translatées (figure 20). Donc, si on a équivariance, alors

$$\Phi(x) = \sum_{g\in G} g.[\tilde{\phi}(x)] = \sum_{g\in G} \tilde{\phi}(g.x) \Rightarrow \Phi(g'.x) = \sum_{g\in G} \tilde{\phi}(g.(g'.x)) = \sum_{g\in G} \tilde{\phi}(g.x) = \Phi(x) \tag{55}$$

car faire une somme sur tous les éléments du groupe G permet d'absorber l'action de g' par changement de variable. Donc, finalement on ne fait pas une simple moyenne de l'image d'origine, ce qui éliminerait toute structuration, mais on va créer une collection de "canaux" ($\left\{\widetilde{\phi_k}(x)\right\}_k$) que l'on moyenne individuellement, et cela ne marche que si les canaux sont équivariants par translation (ou plus généralement par action de groupe).

5.7 Des opérations équivariantes

La question qui se pose immédiatement est de savoir **comment créer des canaux équivariants**? Imaginons que $\phi(x)$ soit **linéaire**, si on impose l'**équivariance** par G (ex. par translation, mais cela s'applique aussi bien par rotation, etc) alors $\phi(x)$ **est une convolution** sur G (voir figure 20). D'où cela fait sens d'utiliser des convolutions partout. De plus, il faut utiliser des non-linéarités (car sinon on perd la structuration), et on veut quelles soient équivariantes elles aussi. Le choix se tourne alors vers des **non-linéarités**

38. NDJE: Notons que S. Mallat introduit la covariance, mais je pense que c'est une habitude de langage car l'équivariance qu'il mentionne aussi est sans doute moins connue. En Physique, la covariance est très repandue: ex. la dérivée covariante en Mécanique des Fluides, la covariance des équations en Mécanique Classique et Relativiste, etc. Dans le cas de figure étudié dans le cours c'est de l'équivariance par translation, mais cette notion est plus générale. Voir par exemple https://arxiv.org/pdf/1805.12301.pdf pour *Rotation Equivariance and Invariance in Convolutional Neural Networks* par B. Chidester et al.

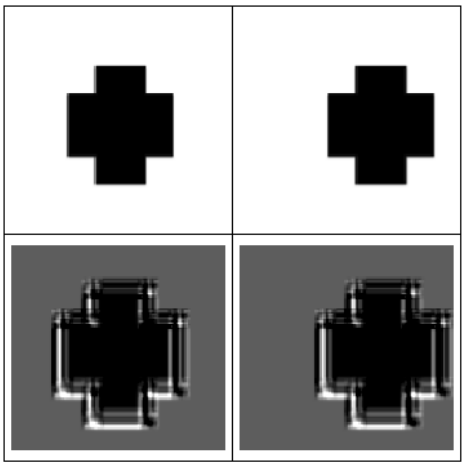

FIGURE 20 – Illustration de la propriété d'équivariance de la convolution par application d'une translation: (haut) de gauche à droite: image origine, image translatée; (bas) de gauche à droite: convolution de l'image d'origine, et convolution de l'image translatée. La convolution de l'image translatée, et la translatée de la convolution de l'image d'origine. Si x est l'image, f la convolution et g la translation alors $f(g.x) = g.f(x)$.

ponctuelles. C'est-à-dire, si on note ρ cette non-linéarité, alors [39]:

$$\rho(x)(u) = \rho(x(u)) \overset{ReLU}{=\!=} \max(0, x(u)) \tag{56}$$

Ce qui veut dire que l'on applique la non-linéarité individuellement à toutes les coordonnées de x, (en pratique on utilise le ReLU ou des sigmoïdes), ce qui conduit à une équivariance de cette opération comme illustré sur la figure 21.

En résumé, on dispose des convolutions et des non-linéarités ponctuelles qui sont toutes des opérations équivariantes, donc il en est de même si on cascade ces opérations. C'est le cas dans un réseau de neurones convolutionel dans les étapes de x à $\Phi(x)$ qui sont donc toutes équivariantes, ensuite on le termine par un pooling pour obtenir un résultat

39. NDJE: je prends un autre notation que S. Mallat pour la différentier des autres transformations, comme les convolutions par exemple utilisées antérieurement.

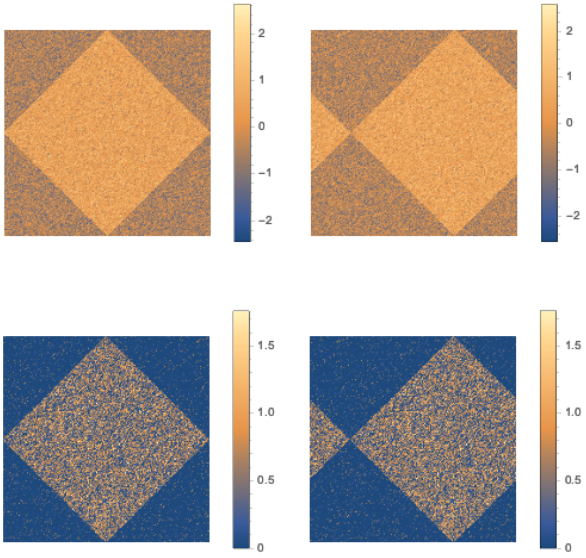

FIGURE 21 – Illustration de la propriété d'équivariance de non-linéarité ponctuelle: (haut) de gauche à droite: image origine, image translatée; (bas) de gauche à droite: rectification (ReLU) de l'image d'origine, et rectification de l'image translatée.

invariant par translation, ce qui peut s'écrire comme ceci:

$$\underbrace{Pool(\Phi(g.x)) = \overbrace{Pool(g.\Phi(x))}^{\text{invariance}}}_{équivariance} = Pool(\Phi(x)) \tag{57}$$

L'information sur la structuration de l'image initiale x est gardée, car on dispose de beaucoup de canaux à l'étape du pooling pour espérer que l'on puisse obtenir une bonne approximation de $f(x)$.

La question qui va nous occuper par la suite est: **comment s'assurer que le résultat doit linéariser les déformations**? Cette contrainte engendre des propriétés spécifiques sur les convolutions que l'on peut employer dans un CNN, et l'on va tomber sur la structuration en multi-échelles: cf. des cascades de "convolution/sous-échantillonnage". Avant d'aborder le sujet, voyons comment la sparcité permet de réduire la dimension du problème.

5.8 La parcimonie (sparsité)

C'est une des trois propriétés qui permet la réduction de dimension avec la séparabilité, symétrie (rappel: "les trois S"). On aimerait que l'image $x \in \mathbb{R}^d$ (avec d grand) soit parcimonieuse, c'est-à-dire qu'on aimerait pouvoir réduire le nombre de variables, pour n'en retenir qu'un petit nombre non nulles. On peut faire cela à travers un opérateur linéaire D, une liste de *dictionnaires*, tel que:

$$D(x) = \begin{pmatrix} \dots \\ \ell_i \\ \dots \end{pmatrix}_{d' \times d} \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}_{d \times 1} = \begin{pmatrix} \dots \\ \langle \ell_i, x \rangle \\ \dots \end{pmatrix}_{d' \times 1} \tag{58}$$

Chaque ligne de D peut être vu comme un *pattern* élémentaire, et donc on projette x sur chacun de ces patterns élémentaires. Le but du jeu est de trouver un jeu de patterns pour que la quasi totalité des produits scalaires soient nuls. Par exemple, on peut penser à utiliser un algorithme de **compression d'image** qui produit beaucoup de coefficients presque nuls, mais **pour produire de la parcimonie, il faut y adjoindre un seuillage.** On utilise par exemple la fonction ReLU avec un seuil $\lambda > 0$:

$$\rho_\lambda(x) \equiv \rho(x - \lambda) = \begin{cases} 0 & x < \lambda \\ x - \lambda & x \geq \lambda \end{cases} \tag{59}$$

Donc, si on applique cette fonction à la sortie de la projection sur les patterns, il s'en suit un seuillage de chaque projection qui produit d'autant plus de coefficients nuls que le seuil est élevé. Par ce biais, on peut éliminer les structures qui ne corrèlent pas avec les patterns du dictionnaires.

Maintenant, on se fixe des dictionnaires **inversibles** (pseudo-) tel que $D^+Dx = x$, alors si naturellement on a $f(x) = f(D^+Dx)$, ce que l'on espère c'est que l'opération de seuillage (débruitage) permette la restitution du signal:

$$\underbrace{f(x)}_{\text{vrai par déf.}} = \overbrace{f(D^+Dx) = f(D^+\rho_\lambda(Dx))}^{\text{débruitage}} \tag{60}$$

Par cette méthode, on veut par exemple faire de la reconnaissance faciale dans une scène: les dictionnaires sont alors tous les types de visages possibles (un peu brutale comme méthode) et calculer les produits scalaires permet de ne garder que les visages pertinents, et par la suite on se demande si on reconnait telle ou telle personne.

L'intérêt du seuillage est qu'il ne reste que $p \ll d'$ coefficients non nuls (nb. on ne sait pas a priori lesquels), ce qui implique que le vecteur $\rho_\lambda(Dx)$ soit de basse dimension. Alors, on peut le projeter avec une **matrice aléatoire W**, soit $\mathbf{W}\rho_\lambda(Dx)$ de dimension $O(p \log d/p)$ qui permet de récupérer le vecteur d'origine. Donc, à un facteur log près on a réduit la dimension du problème à p. Ainsi, **rendre parcimonieux x permet de réduire considérablement la dimensionalité du problème**.

Cependant, des questions se posent: quel est ce dictionnaire (ou ces dictionnaires si on les cascades)? Faut-il les apprendre? On va voir que l'on peut commencer par utiliser des dictionnaires *a priori* qui vont sparsifier le problème, simplement parce qu'on **connait la variable** u. Et donc on a bien de l'information *a priori* pour construire un premier type de dictionnaire (analyse de Fourier à fenêtre, analyse en ondelettes). Et typiquement, ce sont les premiers filtres que l'on met en évidence dans les réseaux de neurones.

Donc, en définitive rien qu'en connaissant la variable d'indexation de x, cf. le u, on peut mettre en œuvre de l'information *a priori* via les trois S: séparabilité, symétrie et sparsité.

6. Séance du 26 Février

NDJE: j'inclus ici les 5 dernières minutes du cours de S. Mallat du 12 fev. car il introduit le sujet de cette leçon: que faire "sans apprendre"?

6.1 Introduction

A grands traits avant 2010, à partir de x on construisait un vecteur de features $\Phi(x)$ avec un peu de math, et ensuite on faisait une régression linéaire. On va voir pourquoi cela fait du sens. On va travailler sur **la variable d'indexation** u, et mettre en œuvre les 3S (séparabilité, symétries et sparsité/parcimonie) pour réduire la dimensionalité du

problème. Ce faisant, on va arriver à la Transformée de Fourier (TF). Mais, pour autant, personne n'utilise la TF pour faire de la reconnaissance de forme. Pourquoi? Primo, elle est instable par déformation, secundo elle n'est pas du tout parcimonieuse: la quasi totalité des coefficients sont non nuls. Donc, il faut se sortir de Fourier, et aller chercher du coté des représentations temps-fréquence ou espace-fréquence pour faire apparaitre la séparabilité et la parcimonie de la plupart des signaux (sons, images). Parmi ces représentations duales, on étudiera la Transformée en Ondelettes (TO), et par la suite on enchainera sur les cascades de filtres. On va s'apercevoir que les CNN font ce genre de choses, au moins dans les premières étapes et que par la suite ils font des choses plus subtiles.

6.2 Équivariance (covariance)

Soit donc l'entrée $x(u)$ et un opérateur L tel que la transformation se note $Lx(u)$. On aimerait que si u est translaté alors la transformée le soit aussi : c'est l'équivariance ou covariance [40]. **Si on impose que L soit linéaire, ce qui est le cas par exemple dans les réseaux de neurones (cf. filtres), alors L est un opérateur de convolution**. Dans la suite, pour simplifier les démonstrations, on considérera que u est continue [41], c'est-à-dire que $u \in \mathbb{R}^q$ (cf. $u \in \mathbb{R}$ pour un son, $u \in \mathbb{R}^2$ pour une image, etc).

Théorème 2. *Soit l'opérateur L linéaire, continue faiblement, équivariant par translation, alors*

$$\exists h \ / \quad Lx(u) = (x * h)(u) = \int x(u-v)h(v)dv \tag{61}$$

Démonstration 2. $x(u)$ peut être représenté comme une intégrale impliquant des Dirac δ, selon

$$x(u) = \int x(v)\delta(u-v)dv \tag{62}$$

En appliquant l'opérateur L agissant sur u, et en introduisant la **réponse impulsionnelle**

40. NDJE: par la suite on utilise les deux mots pour la même notion décrite Sec. 5.6.
41. NDJE: voir le Cours 2018 Sec. 5.2 pour un développement de l'analyse de Fourier en discret.

h définie par $L[\delta(u)] \equiv h(u)$, il vient alors

$$\begin{aligned} L[x(u)] &= \int L[x(v)\delta(u-v)]dv = \int x(v)L[\delta(u-v)]dv \\ &= \int x(v)h(u-v)dv = (x*h)(u) = (h*x)(u) \end{aligned} \tag{63}$$

■

Ce résultat n'est pas spécifique aux translations, c'est vrai pour n'importe quel groupe: si un **opérateur est équivariant** par action du groupe et qu'il est **linéaire**, alors l'opérateur est une **convolution sur le groupe.**

6.3 La Transformée de Fourier

Donc, une fois que l'on a définit l'opérateur de convolution, l'idée est de savoir si on peut le diagonaliser, ce qui nous amène naturellement à la Transformée de Fourier. Si l'on soumet une sinusoïde $e^{i\omega u}$ à l'opérateur, il vient

$$L\left[e^{i\omega u}\right] = \int e^{i\omega(u-v)}h(v)dv = e^{i\omega u}\int h(v)e^{-i\omega v}dv = e^{i\omega u}\hat{h}(\omega) \tag{64}$$

Donc, primo $e^{i\omega u}$ **est un vecteur propre de** L, et **la valeur propre est reliée à sa réponse impulsionnelle** à savoir $\hat{h}(\omega)$ qui n'est autre que la **fonction de transfert de l'opérateur** L [42].

On va revoir (sans démonstration) les propriétés de base de la TF [43]. La TF est omniprésente en Physique notamment, car elle diagonalise les opérateurs linéaires équivariants par translation en particulier les opérateurs différentiels, ce qui en fait un outil de choix pour la résolutions d'équations différentielles.

42. En dimension supérieure à 1, il faut considérer que ωu est un produit scalaire $\omega.u$ avec $(u,\omega) \in \mathbb{R}^q$.
43. S. Mallat indique que sur le site web associé au cours, il y a des liens vers des références/livres sur non seulement la TF mais aussi l'analyse temps-fréquence en général.

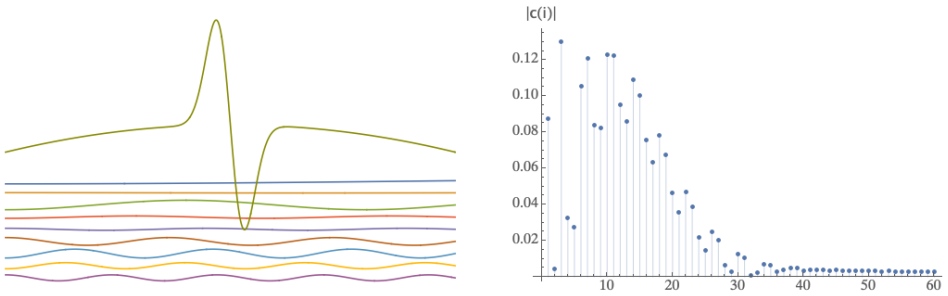

FIGURE 22 – Décomposition d'une fonction en sinusoïdes, ou plutôt selon la Discret Cosine Transform de type III. A gauche les différentes composantes dans l'espace réel, et à droite l'évolution de l'amplitude des coefficients de Fourier.

6.3.1 Inversion

Théorème 3. *Si $x \in \mathbb{R}^p$ est intégrable, c'est-à-dire que $\int |x(u)|du < \infty$, et que sa transformée de Fourier est aussi intégrable, cf. $\int |\hat{x}(\omega)|d\omega < +\infty$, alors*

$$x(u) = \frac{1}{2\pi} \int \hat{x}(\omega) e^{i\omega u} d\omega \tag{65}$$

Donc, ce résultat nous permet de reconstruire un signal à partir de sa transformée de Fourier, à savoir à partir d'une somme de sinusoïdes de pulsations ω dont l'amplitude et phase est $\hat{x}(\omega)$. Comme cela a déjà été indiqué dans le Cours 2018 (Sec 5.2.2), ce résultat n'est pas du tout intuitif: sur la figure 22, la variabilité très rapide locale fait intervenir une sinusoïde à grande fréquence, mais ailleurs les oscillations rapides doivent être compensées exactement pour laisser place à une fonction à variations douces représentées par des sinusoïdes de basses fréquences. Intuitivement, on sent bien que **la décroissance des coefficients** de Fourier, cf. $|\hat{x}(\omega)|$ nous renseigne sur la **régularité** de la fonction $x(u)$. Cependant, la fonction ne présente qu'une irrégularité que **locale** dans l'espace réel, alors que les sinusoïdes sont totalement délocalisées, ce qui est le réel problème de la TF. En fait la décroissance des coefficients de Fourier est donnée par la pire des singularités de la fonction.

On étend le théorème 3 aux fonctions de $L^2(\mathbb{R})$, c'est-à-dire de carré sommable, soit $\int |x(u)|^2 du < +\infty$, voire également l'extension aux distributions.

6.3.2 Quelques propriétés de la Transformée de Fourier

NDJE. Voir le cours de 2018 également sur un point sur des jeux de conventions de la TF.

6.3.2.1 Produit de Convolution

Théorème 4. *La transformée de Fourier d'un produit de convolution est le produit des transformées de Fourier dès lors quelles existent, c'est-à-dire que*

$$\widehat{x * h}(\omega) = \hat{x}(\omega)\hat{h}(\omega) \tag{66}$$

Cela vient directement que les sinusoïdes sont des vecteurs propres de l'opérateur de convolution, et on utilise la propriété d'inversion.

6.3.2.2 Formule de Plancherel

Théorème 5. *Soit x_1 et x_2 des éléments de $L^2(\mathbb{R}^p)$ (cf. fonctions d'énergie finie), le produit scalaire se définit selon*

$$\langle x_1, x_2 \rangle = \int x_1(u) x_2^*(u) du \tag{67}$$

alors on a l'équation suivante[44]

$$\langle x_1, x_2 \rangle = \frac{1}{2\pi} \int \hat{x}_1(\omega) \hat{x}_2^*(\omega) d\omega = \frac{1}{2\pi} \langle \hat{x}_1, \hat{x}_2 \rangle \tag{68}$$

C'est une propriété d'isométrie fondamentale, et une conséquence directe du théorème de convolution.

6.3.2.3 La Dilatation

Soit la fonction $x_s(u)$ qui se définit en dimension p à partir de $x(u)$ selon (figure 23):

$$x_s(u) = \frac{1}{s^p}\, x(\frac{u}{s}) \tag{69}$$

44. NDJE: attention 1) à la convention de la TF et 2) à la dimension p pour obtenir la constante de normalisation de la formule de Plancherel.

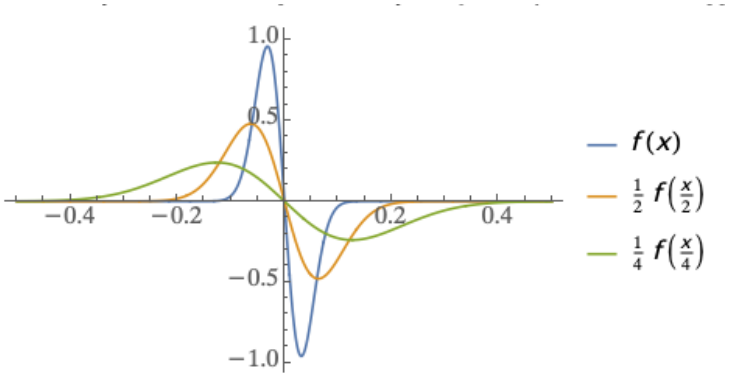

FIGURE 23 – Exemples de dilatations de $f(x)$.

Si $s > 1$ on dilate la fonction, tandis que si $0 < s < 1$ on la contracte. Dans le domaine de Fourier, on obtient aisément que

$$\hat{x}_s(\omega) = \frac{1}{s^{p-1}}\hat{x}(s\omega) \tag{70}$$

c'est-à-dire que si $x(u)$ est dilatée, sa transformée de Fourier est contractée, et vice-versa.

6.3.2.4 Les dérivées

La dérivation se traduit par une multiplication par $i\omega$ ce qui par dérivations successives donne la relation (nb. la fonction $x(u)$ tends vers 0 à l'infini):

$$\hat{x}^{(q)}(\omega) = (i\omega)^q\hat{x}(\omega) \tag{71}$$

Notamment l'analyse des espaces de Sobolev regarde les dérivées non pas dans l'espace réel mais dans celui de Fourier à travers le comportement du produit $\omega^q|\hat{x}(\omega)|$ à l'infini.

6.3.2.5 La Translation

Soit la fonction définie par $x_\tau(u) = x(u-\tau)$, alors sa transformée de Fourier est égale à

$$\hat{x}_\tau(\omega) = e^{-i\omega\tau}\hat{x}(\omega) \tag{72}$$

Quelle est la condition pour qu'un opérateur linéaire soit invariant par translation? Il faut que $e^{-i\omega\tau}$ soit indépendant de τ, c'est-à-dire que $\omega = 0$, or

$$\hat{x}_\tau(0) = \int x(u)du \tag{73}$$

c'est la **moyenne** du signal. C'est un résultat qui se généralise, **si un opérateur linéaire est invariant par l'action d'un groupe, on somme les coefficients sur toutes les orbites pour obtenir la moyenne**. Qu'en est-il dans le non-linéaire?

On a vu que les "recalages" (Sec. 5.5) sont instables par dilatation, donc si on veut un opérateur non-linéaire invariant par translation, il faut utiliser autre chose. La transformée de Fourier (via Plancherel) nous offre un outil de choix, à savoir prenons le module de $\hat{x}$ pour définir la représentation Φ (ici on a une vision avec u une variable discrète) comme le vecteur suivant:

$$\Phi(x) = \{|\hat{x}(\omega)|\}_\omega$$

Or, personne n'utilise la TF pour faire de la reconnaissance de forme. Pourquoi? Une première raison que nous allons voir plus loin est que la TF ne donne pas une représentation parcimonieuse et ne permet pas de zoomer sur des structures, et la seconde que nous avons abordé plus haut, est qu'elle n'est pas stable par déformation locale (cf. les hautes fréquences).

6.3.2.6 Instabilité par déformation

Le problème d'instabilité par déformation nous guide pour définir la bonne représentation et la plus part des contraintes. Une déformation est représentée par l'action du groupe des difféomorphismes selon

$$x(u) \to g.x(u) = x(\theta(u)) = x(u - \tau(u))$$

avec la fonction $\theta \in C^1$ (continument dérivable). Cette dernière fonction est de plus inversible lorsque l'on considère de *petites* déformations (cf. qui ne changent pas la classe de x) qui se traduisent par l'action d'une petite translation locale $\tau \in C^1$. Pour que de plus $\mathbb{I} - \tau$ soit inversible, on impose la contrainte $\|\nabla\tau\|_\infty < 1$ (cf. en dimension 1 cela se traduit par $|\tau'(u)| < 1$).

Prenons pour illustrer cette instabilité de la TF, $\tau(u) = \varepsilon u$ avec $|\tau'(u)| = \varepsilon \ll 1$. Si

$\varepsilon \ll 1$, on ne devrait pas s'attendre à ce que la représentation $\Phi(x)$ soit chamboulée. On veut pour cela que Φ soit différentiable, tout du moins soit assez régulière, ex. comme être Lipschitz (Eq. 46), c'est-à-dire dans le cas présent cela revient à:

$$\|\Phi(g.x) - \Phi(x)\| \leq C\|\Phi(x)\| \underbrace{\|\nabla\tau\|_\infty}_{\text{déformation}} = C\|\Phi(x)\|\varepsilon$$

Autrement dit, on veut que $\|\Phi(g.x) - \Phi(x)\|$ soit du même ordre de grandeur que la taille de la déformation ε. Voyons ce qu'il se passe en Fourier car notre représentation est $\Phi(x) = |\hat{x}(\omega)|$. Or [45]

$$g.x = x[u - \tau(u)] = x[(1-\varepsilon)u] \xrightarrow{TF} \hat{x}[(1+\varepsilon)\omega] \tag{74}$$

Si le signal a de la puissance à hautes fréquences comme deux "bumps" centrés sur $\omega = \pm\omega_0$, alors par l'action de τ la puissance va se concentrer autour de ω_0' telle que $(1+\varepsilon)\omega_0' = \pm\omega_0$ ou $\omega_0' = \pm(1-\varepsilon)\omega_0$ (voir la figure 18 mais cette fois le signe de ε a changé car on est dans le domaine de Fourier). Au bilan, **les bumps se déplacent vers les basses fréquences d'une quantité** $\varepsilon\omega_0$. **Or, rien ne dit que cette quantité soit petite de telle façon que les bumps avant et après transformation ne se superposent pas**. Et donc, au lieu d'avoir $\|\Phi(x) - \Phi(g.x)\| \propto \varepsilon$, on a plutôt [46]

$$\|\Phi(x) - \Phi(g.x)\| = \||\hat{x}(\omega)| - |\hat{x}_\tau(\omega)|\| = 2\||\hat{x}(\omega)|\| = 2||\Phi(x)|| \tag{75}$$

En cela $\Phi(x)$ **ne peut satisfaire la condition de Lipschitz** dès lors qu'il y a de la puissance à hautes fréquences. Or, ces hautes fréquences sont issues de tous les petits détails de la fonction $x(u)$ qui pourtant ne devraient pas nuire à la reconnaissance/classification.

Donc, il faut trouver des représentations invariantes par translation, **stables par déformation** afin de maintenir la différence $\|\Phi(x) - \Phi(g.x)\|$ de l'ordre de l'amplitude de la déformation (au moins au premier ordre). Pour ce faire, **il faut passer par la séparation d'échelles**.

45. en dimension p, il y a un terme multiplicatif $(1-\varepsilon)^{1-p}$ qui est inessentiel pour le propos ici.

46. NDJE: on peut s'en rendre compte en prenant $x(u) = e^{-(1/2)u^2\sigma^2}\cos[u\omega_0]$, la transformée de Fourier correspond à deux gaussiennes de moyennes $\pm\omega_0$ et de sigmas σ. Pour $\omega_0 = 2$ et $\sigma = 0.1$, on voit que pour $\varepsilon = 0.5$ les 2 "bumps" ne se recouvrent plus, et que la différence $\||\hat{x}(\omega)| - |\hat{x}((1+\varepsilon)\omega)|\|$ n'est plus donnée par le développement au premier ordre en ε mais par les 2 séries de "bumps" avant et après la dilatation.

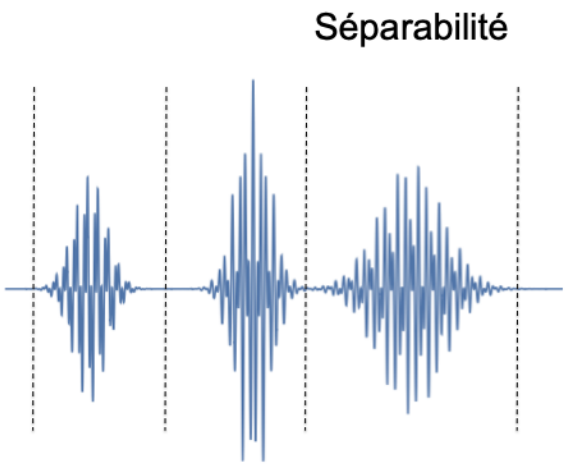

FIGURE 24 – Exemple de l'usage de la séparabilité pour isoler les signaux où il y a de l'information, et sur chaque intervalle on peut procéder à une représentation parcimonieuse (sparcité).

6.4 Représentation temps-fréquence à fenêtre

Les "3S" (séparabilité, symétrie, sparsité) motivent la représentation Temps-Fréquence [47]. Notons que c'est un sujet qui est apparu dans les années 1950 avec Dennis Gabor (1900-1979), physicien hongrois ayant reçu le Prix Nobel en 1971 pour l'invention de l'holographie. Il s'est intéressé aux liens entre Mécanique Quantique et Théorie de l'information, en particulier comment représenter de l'information par exemple quand elle est véhiculée par les sons, la parole. Le problème c'est que pour ce qui concerne la parole, les phonèmes sont oscillants ce qui motiverait la TF pour faire une analyse fréquentielle, mais ils sont aussi **localisés en temps**. Donc, on a besoin d'une analyse combinée Temps-Fréquence. Pour ce faire, on va utiliser la **séparabilité** (Fig. 24) pour isoler les intervalles de temps où il y a du signal, en extrayant sur chaque intervalle de l'information en plus basse dimension, et on va utiliser une représentation où la plupart des coefficients sont quasiment nuls (**sparsité/parcimonie**).

47. Il s'agit plus généralement, tout autre couple de variables duales par TF.

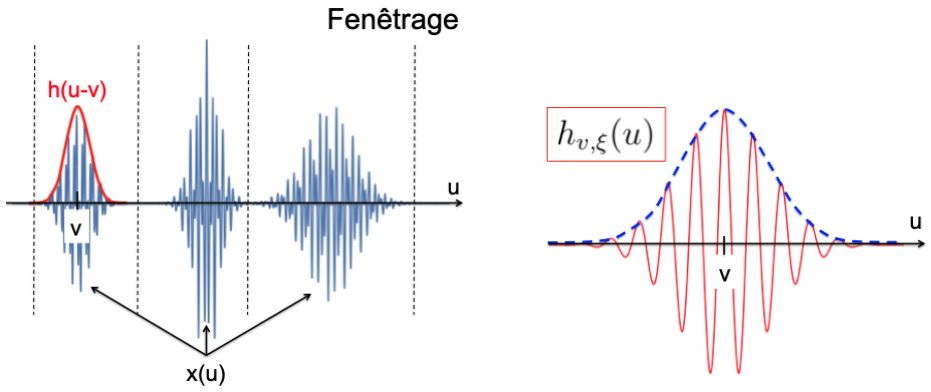

FIGURE 25 – (gauche): Exemple de l'utilisation d'une fenêtre h centrée sur $u = v$ afin d'isoler une partie de la trame $x(u)$. La généralisation en 2D est immédiate par une fenêtre gaussienne bidimensionnelle. (droite): représentation de la fonction $h_{v,\xi}$.

6.4.1 Le fenêtrage (Short Time Fourier Transform)

L'idée de l'analyse temps-fréquence va consister à reprendre l'intégrale de Fourier, et comme on veut introduire de la localisation (séparabilité), on va multiplier $x(u)$ par une **fenêtre** (Fig. 25 gauche) h pour isoler une tranche de signal, et on applique ensuite une TF pour en tirer les fréquences principales. Ainsi, on définit la **Short Time Fourier Transform**, selon

$$\boxed{Sx(v,\xi) = \int x(u)h(u-v)e^{-i\xi u}du = \int x(u)h^*_{v,\xi}(u)du = \langle x, h_{v,\xi}\rangle} \tag{76}$$

C'est une extension de la TF où l'on regarde simultanément la variable temporelle v et la variable de fréquence ξ. Quelles sont les propriétés de ce type de transformation, et comment choisir la fenêtre h?

On peut voir la transformation comme une corrélation du signal $x(u)$ avec une function $h_{v,\xi}(u)$ qui n'est autre qu'une fenêtre centrée en v multipliée par une sinusoïde (Fig. 25 droite). Dans ce contexte, **la taille de la fenêtre est fixe**, les seuls paramètres que l'on peut faire varier sont: la position centrale v et la fréquence ξ. Dans la littérature $h_{v,\xi}(u)$ est appelé *atome temps-fréquence*, à cause du lien avec la Mécanique Quantique (cf. le paquet d'ondes).

Maintenant, en utilisant la formule de Plancherel, on peut avoir une vision fréquentielle de l'équation 76:

$$Sx(v,\xi) = \frac{1}{2\pi}\int \hat{x}(\omega)\hat{h}^*_{v,\xi}(\omega)d\omega = \frac{1}{2\pi}\langle \hat{x}, \hat{h}_{v,\xi}\rangle \tag{77}$$

Ainsi, $Sx(v,\xi)$ **est aussi la corrélation entre la TF de** $x(u)$ **et la TF d'un atome**. Quelle est la forme de cette dernière? En utilisant 1) que $h_{v,\xi}(u) = \tilde{h}(u-v)e^{i\xi v}$ avec $\tilde{h}(u) = h(u)e^{i\xi u}$, 2) la règle qu'il y a un lien entre translation et multiplication par une phase alors $\hat{\tilde{h}}(\omega) = \hat{h}(\omega-\xi)$ et donc

$$\boxed{h_{v,\xi}(u) = h(u-v)e^{i\xi u}} = \tilde{h}(u-v)e^{i\xi v} \xrightarrow{TF} \boxed{\hat{h}_{v,\xi}(\omega) = e^{-i(\omega-\xi)v}\hat{h}(\omega-\xi)} \tag{78}$$

La question est de savoir quelle fenêtre choisir? On aimerait bien à la fois faire une bonne localisation autour de $u = v$ et de $\omega = \xi$. Peut-on définir une boîte aussi petite que l'on veut? La réponse est non comme on peut le constater sur la figure 27. **Le principe d'incertitude d'Heisenberg est à l'œuvre pour contraindre la taille de la boite**. On va imposer que $||h||^2 = 1$, c'est-à-dire que l'intégrale de son carré est égale à 1 (nb. c'est le cas pour l'expression de la gaussienne Eq. 92). Définissons la largeur en temps et en fréquence. Pour la partie temporelle et fréquentielle, on peut se définir des variances avec les masses correspondantes:

$$\begin{aligned}
\sigma_u^2 &\equiv \int (u-v)^2 \left\|h_{v,\xi}(u)\right\|^2 du \\
&= \int (u-v)^2 \left|h(u-v)\right|^2 du = \int u^2 h^2(u)du \qquad \left(\text{nb. } \int h^2(u)du \equiv 1\right) \qquad (79)\\
\sigma_\omega^2 &\equiv \frac{1}{2\pi}\int (\omega-\xi)^2 \left\|\hat{h}_{v,\xi}(\omega)\right\|^2 d\omega = \frac{1}{2\pi}\int \omega^2 |\hat{h}(\omega)|^2 d\omega \qquad (80)
\end{aligned}$$

On remarque que σ_u et σ_ω ne dépendent pas du lieu (v,ξ) dans le plan (u,ω). Mais, peut-on fixer arbitrairement σ_u et σ_ω aussi petit que l'on veut? Prenons des exemples simples:

- Si on prend un Dirac centré sur $u = v$, donc localisation à l'extrême dans le temps, mais en Fourier le Dirac se transforme en la fonction 1, donc totalement délocalisée en fréquence;
- Symétriquement, une sinusoïde est localisée sur 1 fréquence, c'est un Dirac en fréquence, et par TF Inverse, elle se transforme en 1 sur l'échelle temporelle, donc totalement délocalisée.

Donc squeezer à l'extrême dans une dimension, ne marche pas. Prenons une fenêtre $h(u)$ et opérons une dilatation s, par TF cela donne:

$$\frac{1}{s}h(u/s) \xrightarrow{TF} \hat{h}(s\omega) \tag{81}$$

Donc, si on localise dans le temps Δu petit (scaling en s), alors la largeur en fréquence augmente $\Delta\omega$ par une scaling en $1/s$.

Cependant, voyons le problème sous un autre angle: celui des dérivées. Pour préciser qu'une fonction est régulière, on dit en général quelle possèdent des dérivées, et que ces dernières sont d'énergie finie (cf. de carré intégrable): $\int \left|h^{(p)}(u)\right|^2 du < \infty$. Or, en utilisant l'égalité de Parseval, on va avoir:

$$\int \left|h^{(p)}(u)\right|^2 du = \frac{1}{2\pi}\int \left|\widehat{h^{(p)}}(\omega)\right|^2 d\omega \tag{82}$$

$$= \frac{1}{2\pi}\int \omega^{2p}\left|\hat{h}(\omega)\right|^2 d\omega \tag{83}$$

Donc la régularité de la fonction se traduit dans le domaine de Fourier par

$$\boxed{\int \left|h^{(p)}(u)\right|^2 du < +\infty \Leftrightarrow \int \omega^{2p}\left|\hat{h}(\omega)\right|^2 d\omega < +\infty} \tag{84}$$

Donc, **plus la fonction est régulière, plus la décroissance des "coefficients" (ou de la TF) de Fourier doit être rapide**. Ainsi, si on essaye de concentrer une fenêtre en temps, on crée forcément une fonction irrégulière, donc ses coefficients de Fourier ne peuvent décroitre rapidement, d'où la délocalisation en fréquence.

Par contre, on sent que la surface de la boîte semble une constante comme on peut s'en rendre compte également sur la figure 27 lorsque ξ varie. Le théorème d'incertitude traduit tous les effets de squeezer l'une ou l'autre fenêtre $h_{v,\xi}(u)$ et $\hat{h}_{v,\xi}(\omega)$ en donnant une borne sur la surface de localisation temps-fréquence.

Théorème 6. *Soit la fenêtre (ici la partie non-oscillante) $h \in L^2(\mathbb{R})$ telle que $\|h\|^2 = 1$ (énergie finie), centrée en 0, c'est-à-dire $\int u|h(u)|^2du = \int \omega|\hat{h}(\omega)|^2d\omega = 0$. Par exemple, on peut penser à la gaussienne Eq. 92. Alors si on définit les variances en temps et*

fréquence comme suit

$$\sigma_u^2 = \int u^2 |h(u)|^2 du \qquad \sigma_\xi^2 = \frac{1}{2\pi} \int \omega^2 |\hat{h}(\omega)|^2 d\omega \tag{85}$$

alors

$$\boxed{\sigma_u \sigma_\omega \geq \frac{1}{2}} \tag{86}$$

Ce théorème est l'expression du **Principe d'incertitude** en Mécanique Quantique où la fonction d'onde (ex. les $h_{v,\xi}(u)$) au carré donne la probabilité de présence d'un électron au voisinage d'un certain point: on ne peut à la fois mesurer avec une infinie précision la position et la quantité de mouvement qui sont deux quantités duales par TF. Ce résultat est une conséquence des propriétés de régularité de la fonction de localisation.

Démonstration 6. Prenons donc le produit $\sigma_u^2 \sigma_\omega^2$, il vient

$$\sigma_u^2 \sigma_\omega^2 = \left(\int u^2 |h(u)|^2 du \right) \times \left(\frac{1}{2\pi} \int \omega^2 |\hat{h}(\omega)|^2 d\omega \right) \tag{87}$$

Remarquons que $\widehat{h'(u)}(\omega) = i\omega \hat{h}(\omega)$ et en utilisant Plancherel on a

$$\frac{1}{2\pi} \int d\omega \ \omega^2 |\hat{h}(\omega)|^2 = \frac{1}{2\pi} \int d\omega \ |\widehat{h'(u)}(\omega)|^2 = \int du \ |h'(u)|^2 \tag{88}$$

puis en appliquant l'inégalité de Cauchy-Schwarz, il vient

$$\sigma_u^2 \sigma_\omega^2 = \left(\int u^2 |h(u)|^2 du \right) \times \left(\int |h'(u)|^2 du \right) \geq \left(\int |u \ h^*(u) h'(u)| \, du \right)^2 \tag{89}$$

Or, remarquons que

$$|u \ h^*(u) h'(u)| \geq \frac{u}{2} \left(h^*(u) h'(u) + h(u) h'^*(u) \right) = \frac{u}{2} \times \frac{d|h(u)|^2}{du} \tag{90}$$

et utilisant une intégration par parties avec la contrainte [48] $u|h(u)|^2 \xrightarrow[u \to \pm\infty]{} = 0$ alors

$$\sigma_u^2 \sigma_\omega^2 \geq \frac{1}{4} \left(\int |h(u)|^2 \ du \right) = \frac{1}{4} \tag{91}$$

48. C'est une contrainte qui simplifie la démonstration, mais le résultat est plus général et se passe de cette contrainte.

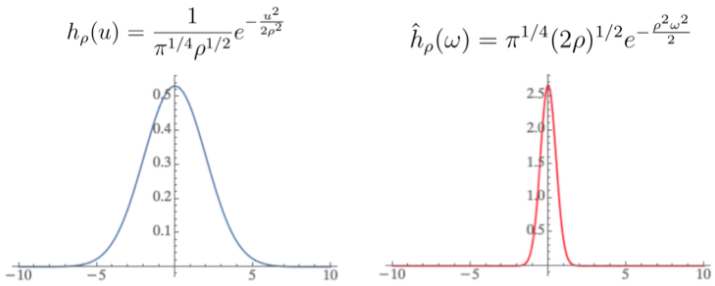

FIGURE 26 – Gaussienne de largeur $\rho = 2$ (à gauche) et sa TF (à droite).

■

Notons que le résultat repose sur le fait que si on veut squeezer $h(u)$ l'intégrale de celle-ci devient très grande et se délocalise donc en Fourier. Dans le cas d'une **fenêtre gaussienne** satisfaisant $\|h\|^2 = 1$ et centrée, on a

$$\boxed{h_\rho(u) = \frac{1}{\pi^{1/4}\rho^{1/2}} e^{-\frac{u^2}{2\rho^2}} \xrightarrow{TF} \hat{h}_\rho(\omega) = \pi^{1/4}(2\rho)^{1/2} e^{-\frac{\rho^2\omega^2}{2}}} \tag{92}$$

et

$$\sigma_u^2 = \frac{\rho^2}{2}, \ \sigma_\omega^2 = \frac{1}{2\rho^2} \Rightarrow \sigma_u^2\sigma_\omega^2 = \frac{1}{4} \tag{93}$$

Donc, dans ce cas la borne inférieure est atteinte, c'est le mieux que l'on puisse faire. La figure 27 représente des ellipses de demi-grands axes $(\sigma_u, \sigma_\omega)$ évoluant en fonction du lieu d'introspection.

6.4.2 Le spectrogramme

Revenons à la définition de la STFT (Eq. 76) avec x translatée $x_\tau(u) = x(u-\tau)$ avec τ très petit par rapport à la fenêtre de $h(u)$ alors

$$\begin{aligned} Sx_\tau(v,\xi) &= \int x(u-\tau)h(u-v)e^{-i\xi u}du = e^{-i\xi\tau}\int x(u)h(u+\tau-v)e^{-i\xi u}du \\ &\simeq e^{-i\xi\tau}Sx(v,\xi) \end{aligned} \tag{94}$$

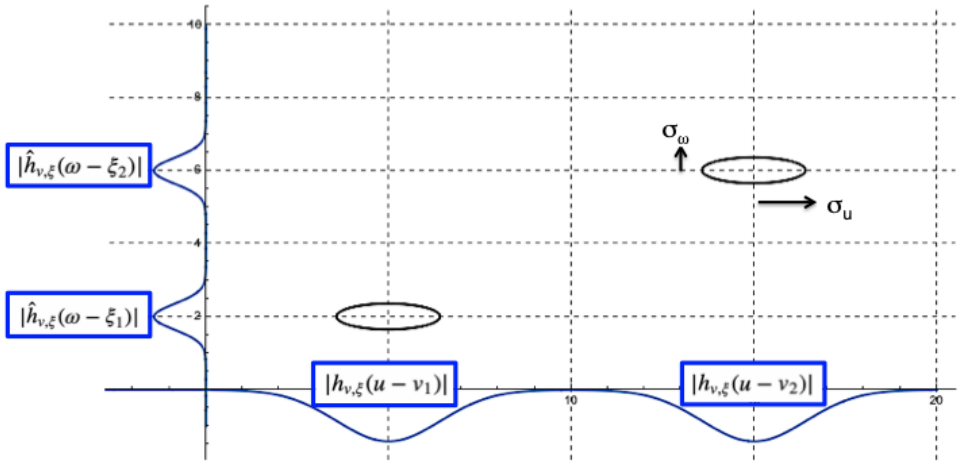

FIGURE 27 – Représentation dans le plan (u, ω) de la localisation de l'introspection par un atome $h_{v,\xi}$ dans le domaine temporel (u), et par sa transformée de Fourier dans le domaine fréquentiel (ω).

Donc, si l'on veut être indifférent aux petites translations, il suffit de prendre le module $|Sx(v, \xi)|$ pour enlever le terme de phase, **c'est le spectrogramme**. Est-ce que cet outil permet de mieux voir les structures? Sur la figure 27, on a représenté la "boîte" d'introspection, ici représentée par une ellipse de demi-grands axes (σ_u, σ_ξ), obtenue par un atome $h_{v,\xi}$, que l'on déplace dans le plan (u, ω).

6.4.3 Quelques exemples

La figure 28 présente la représentation temporelle et le spectrogramme d'un signal composé de 2 chirps, tandis que la figure 29 montre le spectrogramme d'un signal qui a une divergence à $t \approx 1$: c'est un modèle du signal des chauve-souris. Notez qu'au voisinage de la singularité la taille de boite entraine un "lavage" car elle devient trop large en fréquence.

Ce type de phénomène est générique dès lors que l'on a des transitoires: exemple quand on veut détecter l'attaque d'un morceau de musique, on aimerait détecter le moment précis, mais alors le Principe d'incertitude impose que la fenêtre devient grande en fréquence et alors on ne peut distinguer les harmoniques et les instruments. Enfin, la structure d'un

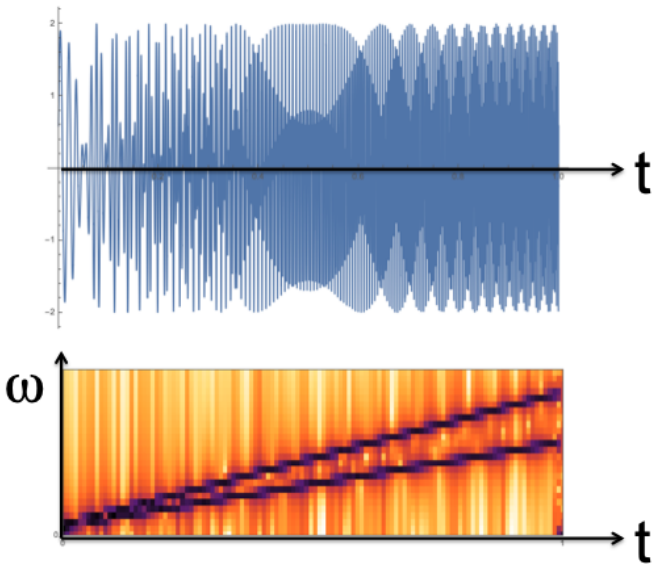

FIGURE 28 – Spectrogramme de la superposition de 2 chirps. Un chirp est un signal du type $x(t) = \sin(\phi_0 + \pi t^2 + 2\pi f_0 t)$. Si la représentation temporelle est complexe, le spectrogramme présente 2 fonctions linéaires croissantes en fréquence.

spectrogramme (Fig. 30) peut aussi être très complexe avec tout un tas de structures que l'on aimerait capturer.

6.4.4 Limitations de la STFT

On voit par l'intermédiaire des spectrogrammes où il y a des singularités/transitoires que le fenêtrage a des limitations. Il y en a un autre qui tient à son instabilité par déformation: en effet, la Short Time Fourier Transform est qu'en même une Transformation de Fourier et partage ce problème. Si deux notes ne différent que par une petite dilatation alors

$$x(u) \to x(u(1-\tau)) \Rightarrow Sx_\tau(v,\xi) \approx Sx(v,(1+\tau)\xi) \tag{95}$$

Donc, si $Sx(v,\xi)$ se concentre en $\xi = 1, 2, 3, \ldots$ alors $Sx_\tau(v,\xi)$ se concentre en $1-\tau, 2-2\tau, 3-3\tau, \ldots$ c'est-à-dire que plus on monte dans les hautes fréquences plus les valeurs des fréquences de concentration de $Sx(v,\xi)$ et $Sx_\tau(v,\xi)$ vont se décaler. Ainsi, il sera de plus en plus difficile de reconnaitre que ces deux sons sont en faites semblables.

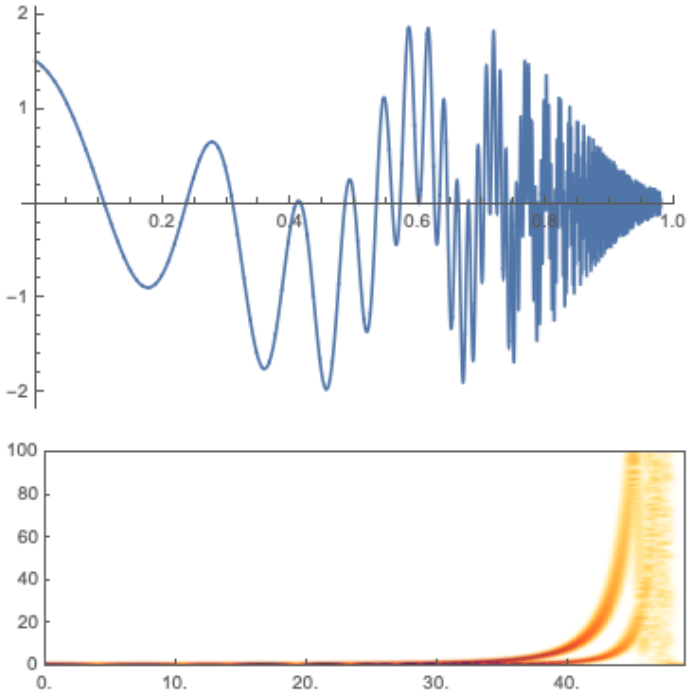

FIGURE 29 – Spectrogramme de la superposition de 2 signaux de type $\cos(a/(t-1)^2)$ avec un damping facteur logistique.

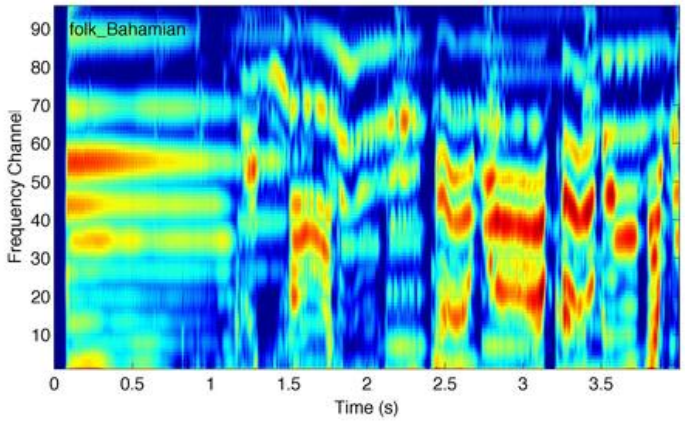

FIGURE 30 – Spectrogramme d'une musique.

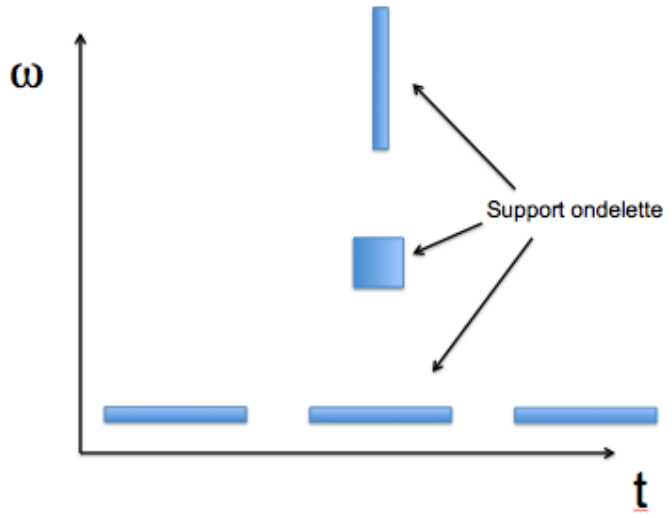

FIGURE 31 – Ajustement de la taille de la boite d'analyse suivant la fréquence.

C'est pour cette raison que l'analyse par fenêtrage n'est pas utilisée en traitement du son, ni d'ailleurs ce n'est pas cela qui est fait par l'oreille. Le truc, va être de changer la taille des boites d'analyse comme par exemple sur la figure 31. On va pratiquer une **analyse multi-résolution par Ondelettes** pour laquelle **la notion d'échelle** apparaît à la place de la notion de fréquence. On va tomber sur des descripteurs (MFCC) qui capturent beaucoup d'invariants (changement d'amplitude, translation en temps, en amplitude), ils sont très compacts et sont basés sur une parcimonie du signal. Jusqu'en 2010, ils dominaient l'analyse du son, puis avec les réseaux ils ont disparus, mais ils réapparaissent car on identifie des filtres communs.

7. Séance du 4 Mars

7.1 Préambule

NDJE: Dans cette séance S. Mallat nous entraine dans le monde des ***Ondelettes****. Vous pouvez également vous reporter à la Sec. 6 du Cours de 2018. Cependant, S. Mallat non seulement va montrer les propriétés classiques des Ondelettes dans l'analyse temps-fréquence, mais aussi prouver que c'est la représentation adaptée pour la linéarisation des*

***déformations**, ce qui est très important pour la séparation d'échelles, pour la reconnaissance. Qui plus est on retrouve dans les CNN les fameuses cascade filtrage/échantillonnage qui produisent ces séparations d'échelles et donc on verra le lien avec les Ondelettes.*

7.2 Temps-Fréquence par Ondelettes

Motivons de trouver une autre représentation que celle de la transformée de Fourier à fenêtre (Sec. 6.4.1):

- on veut pouvoir **isoler les transitoires de façon très précise**, donc ne pas délocaliser l'information temporelle (surtout à hautes fréquences);
- on veut être **stable par déformation**: c'est-à-dire qu'une petite déformation doit se traduire également par une petite différence dans la représentation et non le phénomène décrit Sec. 6.4.4;
- enfin le Principe d'incertitude nous donne des limitations sur la taille de la boite d'introspection (Th. 6), mais on veut pouvoir **capturer à la fois les petites échelles et les grandes échelles** de temps. Par exemple, pensez en musique aux échelles de temps d'une note, d'un accord, d'une mélodie, le mouvement, etc. Donc il y a des structures à toutes les échelles et celles des longues durées est un effort de la recherche actuelle.

7.2.1 La famille d'ondelettes

Soit une fonction $\psi(u) = h(u)e^{i\xi u}$ où $h(u)$ est une fenêtre (voir par exemple la figure 25). Cependant, par rapport à la STFT (Eq. 76), **on fixe la fréquence** ξ. On va opérer 2 transformations: **une translation** ($b \in \mathbb{R}$) **et une dilatation** ($s \in \mathbb{R}^{+*}$) que l'on applique à l'ondelette de base $\psi(u)$ pour donner la famille

$$\boxed{\psi_{s,b}(u) \equiv \frac{1}{s}\psi\left(\frac{u-b}{s}\right) \equiv \psi_s(u-b)} \tag{96}$$

NDJE: Pour des lecteurs des notes du cours de 2018 ou du livre de S. Mallat, ils auront remarqué le changement de normalisation, cf. le passage d'un $1/\sqrt{s}$ à $1/s$. J'en ai discuté avec lui et il me semble intéressant de retracer la motivation de ce changement de normalisation. Dans l'esprit de 2018, on était dans la vision "produit scalaire" de la transformée

en Ondelettes avec dans l'idée de construire une base orthonormale avec un échantillonnage en échelle et en espace. Dans ce contexte la normalisation en $1/\sqrt{s}$ s'impose tout comme par la suite l'apparition du complexe conjugué dans la définition du coefficient en Ondelette $Wx(v,s)$. Dans ce cours de 2020, S. Mallat veut insister sur la vision "filtrage convolutionnel" de la transformation en Ondelettes, dès lors on peut vouloir définir $Wx(v,s)$ comme un produit de convolution et simplifier également le calcul dans l'espace de Fourier, d'où le changement de normalisation en $1/s$ en 1D, et par la suite $1/s^2$ en 2D. Cette mise au point sur d'aucun dirait "une simple normalisation" permet de bien cerner la philosophie du point de vue de cette année.

7.2.2 La transformée en Ondelettes

On définit la Transformée en Ondelettes (TO) de la fonction $x(u)$ par l'ondelette ψ_s selon (sauf mention du contraire les bornes d'intégrations sont $\pm\infty$)

$$\boxed{Wx(v,s) \equiv (x * \psi_s)(v) = \int x(u)\psi_s(v-u)du = \int x(v-u)\psi_s(u)du} \tag{97}$$

Donc, **la transformée en ondelettes est vue comme la convolution** de x par l'ondelette ψ_s.

En utilisant les règles sur la transformée de Fourier (Tab. 1) on a

$$\widehat{\psi_s}(\omega) = \hat{\psi}(s\omega) \qquad \text{et} \qquad \widehat{\psi_s(v-u)}(\omega) = e^{i\omega v}\widehat{\psi_s}(\omega) \tag{98}$$

Si maintenant on applique le théorème de Parseval, alors l'intégrale donnant $Wx(v,s)$ est égale à

$$\boxed{Wx(v,s) = \frac{1}{2\pi}\int \hat{x}(\omega)\hat{\psi}(s\omega)e^{i\omega v}d\omega} \tag{99}$$

qui indique que l'intégrale a des contributions dans le domaine fréquentielle où $\hat{\psi}(s\omega)e^{i\omega v}$ est non nul. Comme

$$\hat{\psi}(\omega) = \hat{h}(\omega - \xi) \tag{100}$$

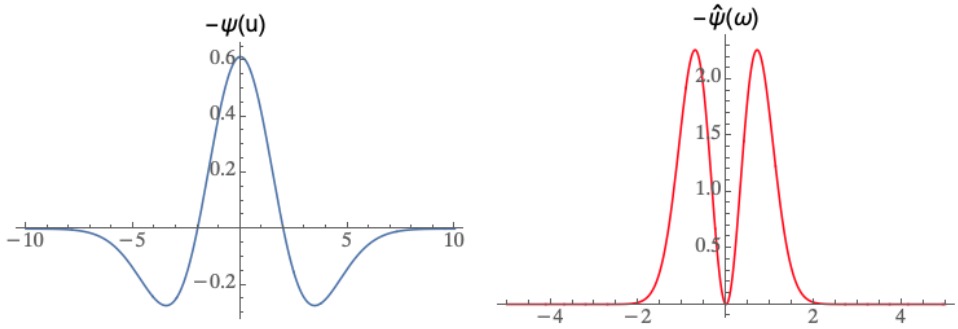

FIGURE 32 – Exemple d'une ondelette en forme de chapeau mexicain (Eq. 102), ainsi que sa TF (Eq. 103) avec $\sigma = 2$. Nb. c'est l'ondelette par défaut de Mathematica.

on va imposer une condition sur l'ondelette telle que

$$\boxed{\int_{-\infty}^{+\infty} \psi(u)du = \hat{\psi}(0) = \hat{h}(-\xi) = 0} \tag{101}$$

On peut avoir affaire soit à des ondelettes *complexes* (analytiques) qui sont plutôt utilisées pour étudier l'évolution de signaux à fréquence bien définie, soit à des ondelettes *réelles* pour des signaux à transitions temporelles rapides. *NDJE: Dans la suite, nous allons considérer les ondelettes réelles, et une section mentionnera les différences dans le cas des ondelettes analytiques qui n'ont pas pu être traitées par S. Mallat faute de temps.*

Voici un exemple de ψ réelle qui satisfait la condition Eq. 101, elle a la forme d'un chapeau mexicain:

$$\psi(u;\sigma) \equiv \frac{2\left(\frac{u^2}{\sigma^2} - 1\right) e^{-\frac{u^2}{2\sigma^2}}}{\pi^{1/4}\sqrt{3\sigma}} \tag{102}$$

Sa transformée de Fourier est alors donnée par

$$\hat{\psi}(\omega;\sigma) = -2\sqrt{\frac{2}{3}}\pi^{1/4}\sigma^{5/2}\omega^2 e^{-\frac{1}{2}\sigma^2\omega^2} \tag{103}$$

La figure 32 donne les graphes de ψ et $\hat{\psi}$ pour $\sigma = 2$. Selon la valeur de s (dilatation/contraction) la position du maximum ainsi que la largeur de l'ondelette en fréquence changent, tout comme la position du premier zéro en temps. Dans le cas de l'ondelette

en chapeau mexicain que l'on dilate, la position du 1er zéro en temps est donné par $u_s^* = s\sigma \propto s$ et celle du maximum en fréquence se trouve en $\omega_s^* = (\sqrt{2}/\sigma)/s \propto 1/s$. Quand s diminue (resp. augmente) la position du maximum en fréquence se décale vers les hautes (resp. basses) fréquences. C'est l'inverse pour la position du premier zéro en temps. Ceci est illustré sur la figure 33.

Comparativement à la transformée de Fourier en fenêtre (Fig.27), **la taille de la fenêtre d'introspection change dans le plan temps-fréquence**, comme on peut le constater sur la figure 34. Mais, **la surface de la boite ne change pas** à cause du principe d'incertitude. Donc, ce n'est pas tant que l'on aurait une meilleure résolution avec une TO, mais **à basse fréquence, là où le signal évolue peu, la taille en u est grande, par contre à hautes fréquences la boîte s'ajuste pour obtenir une meilleure localisation temporelle**. Enfin, **par translation selon u, à ω fixée, la taille de la boite ne change pas**.

7.2.3 Quelques exemples

Sur la figure 35, le signal $x(u) = \sin(2000\pi u^2)$ est analysé à la fois par une TF à fenêtre (STFT) et une TO dont la taille de la fenêtre suit le schéma de la figure 34. Pour la STFT, la boite d'analyse est de taille constante, ainsi la délocalisation temps-fréquence est constante en fonction de la fréquence et du temps. Pour la TO, l'effet de l'agrandissement de la boite le long de l'axe des fréquence au fur et à mesure que la fréquence augmente se voit nettement, il domine complètement de scalogramme [49]. Pour ce type de signal qui ne présente pas de singularité, la STFT est plus adaptée car elle est plus parcimonieuse, à savoir il y a moins de coefficients au delà d'un seuil (ex. 0.5).

Le phénomène de perte de résolution fréquentiel à hautes fréquence de la TO est encore plus gênant quand les deux signaux se ressemblent (voir la figure 36). L'ondelette est trop large en fréquence, et des phénomènes de battements entre les deux signaux sont visibles: on est plus capable de les séparer. Donc, **ici aussi la TO n'est pas utile car il n'y a pas de phénomènes transitoires**.

NDJE: *Ensuite pour les phénomènes transitoires, il vaut mieux se référer à la vidéo du cours car S. Mallat fait écouter les effets sonores.*

49. NDJE: le mot *scalogramme* est le terme consacré pour l'analyse temps-fréquence en TO.

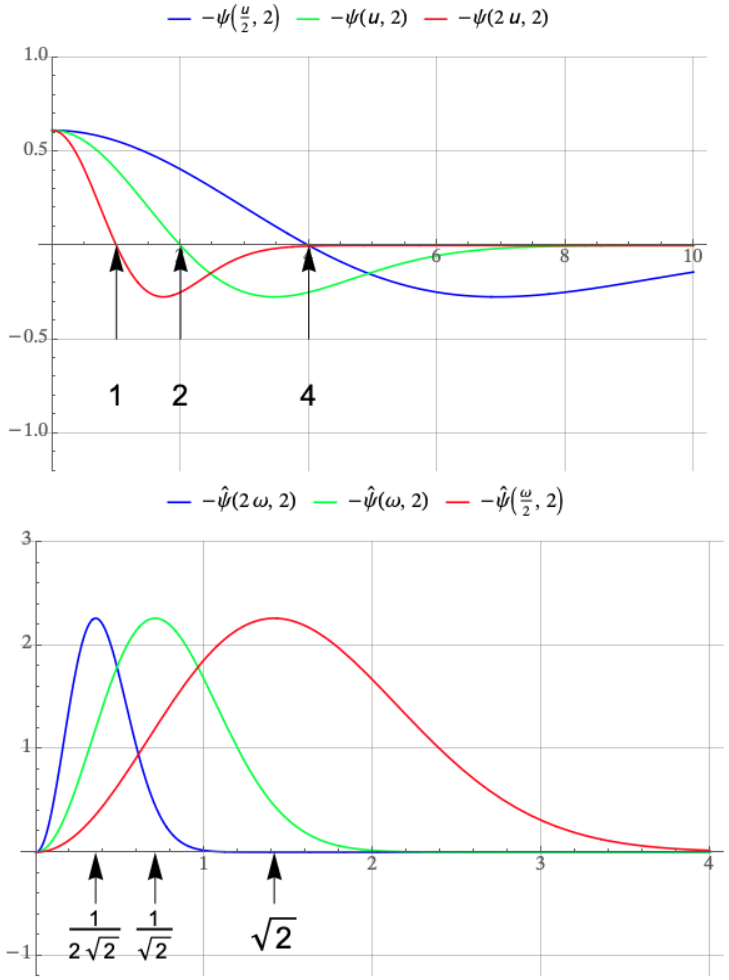

FIGURE 33 – Exemple d'évolution de $\psi_s(u) \propto \psi(u/s)$ et de $\hat{\psi}_s(\omega) \propto \hat{\psi}(s\omega)$ en fonction de $s = 2, 1, 1/2$ (cf courbes bleu, vert, rouge). La position su premier zéro de ψ_s suit une loi en s alors que la position du maximum de $\hat{\psi}_s$ suit une loi en $1/s$. Illustration avec l'ondelette en chapeau mexicain de la figure 32.

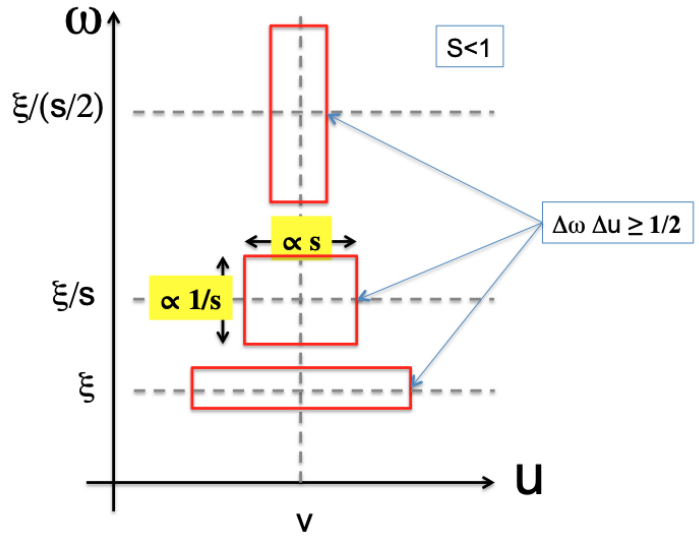

FIGURE 34 – Évolution du support de l'ondelette dans le plan temps-fréquence quand on change le facteur d'échelle.

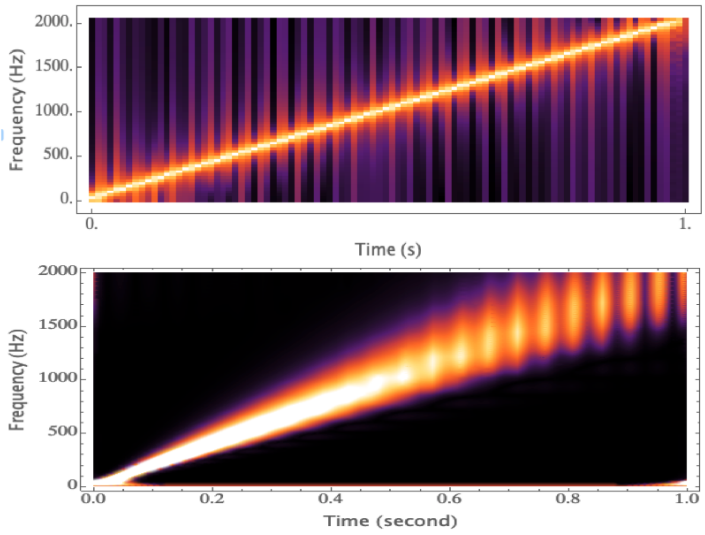

FIGURE 35 – Comparaison entre une analyse par fenêtrage dont la taille de boite d'introspection est fixe (haut), et une analyse par ondelettes (bas) où la taille de boite évolue selon le schéma de la figure 34. Le signal $x(u) = \sin(2000\pi u^2)$ échantillonné au rythme de $1/4095$ sur $u \in [0, 1]$. Pour la TO, on a utilisé une ondelette de Gabor $1/\pi^{1/4} e^{-u^2/2} e^{i\xi u}$ avec $\xi = 6$.

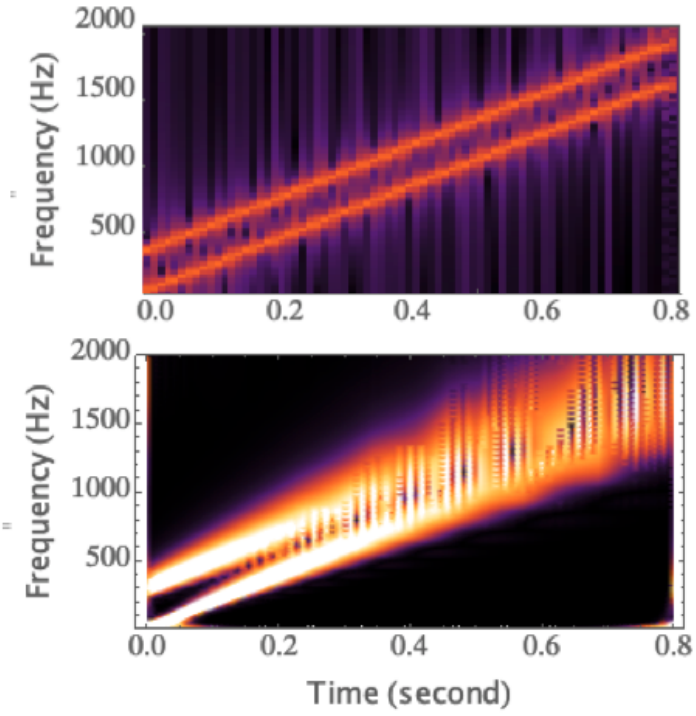

FIGURE 36 – Superposition de 2 signaux du type de celui utilisé pour la figure 35: (haut) par STFT, (bas) par TO.

Maintenant, si le signal est légèrement dilaté alors on peut se demander comment évolue sa transformée en ondelettes. C'est le pendant de l'étude Sec. 6.4.4. Il vient

$$x(u) \to x(u(1-\tau)) \Rightarrow v \to v(1-\tau), s \to s(1-\tau), \xi \to \xi(1+\tau) \tag{104}$$

Donc, les harmoniques du signal se décalent vers les hautes fréquences et comme la taille de boite s'allonge le long de l'axe des fréquences, il s'en suit une confusion des harmoniques. **Mais en contre-partie ce manque de résolution va donner de la stabilité par déformation**. En effet, la STFT est certes précise mais l'instabilité à de petites perturbations (déformations) rend caduque cette qualité car elle tend à faire percevoir différents des signaux qui pourtant sont très semblables. Cela nuit finalement à la classification. La TO plus stable par déformation, au prix d'une moins bonne résolution fréquentielle à hautes fréquences, va être plus adaptée. Donc, **ce n'est pas tant la résolution qui va compter mais la stabilité**.

7.3 Digression naturaliste

Avant d'aborder les points techniques de la transformée en ondelettes, S. Mallat nous fait passer par une description de la psychophysique de l'appareil auditif qui est le sujet d'un séminaire associé (12 Mars 20, Prof. Shihab Shamma). Le schéma de l'oreille humaine (Fig. 37) donne les principaux récepteurs sonores qui transmettent l'onde de pression et la transforment en signaux électriques dans la cochlée. L'organe de Corti situé au centre de la cochlée est divisé en environ 10000 cellules couvertes de cils qui baignent dans un liquide. Suivant la fréquence du signal sonore, il est analysé par une partie spécifique de l'organe de Corti: la partie basse analyse un son aigu de longueur d'onde courte, tandis que la partie haute (apex) analyse un son grave de grande longueur d'onde. **La modélisation de la réponse des cils est la suivante: ils agissent comme un filtre passe-bande qui ressemble très fortement à une ondelette** (sauf à basse fréquence où la largeur est constante). Si on déploie la cochlée, elle se présente comme un long tube de section qui varie en se rétrécissant, et la position le long de ce tube donne en fait le facteur d'échelle de l'ondelette. Qui plus est la représentation est plutôt **une échelle logarithmique** (cf. $\log s$). La question est pourquoi un organe qui a bénéficié de millions d'années d'adaptation ressemble si bien à une transformée en ondelettes? Sans doute que la localisation de transitoires, ainsi que la

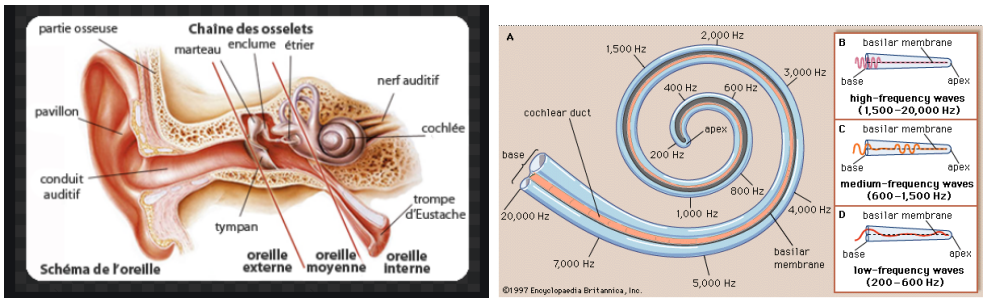

FIGURE 37 – Schéma de l'oreille humaine et de la cochlée.

stabilité aux petites déformations, permettent une meilleure reconnaissance/classification.

Une fois le signal électrique produit, il est transmis à travers le nerf auditif (pour faire simple) au cortex auditif qui comporte deux grandes zones (Fig. 38). En fait,dans la partie A1 du cortex auditif le traitement revient, après le filtrage $x * \psi_s$, à procéder à une rectification par une fonction ρ, comme dans un CNN, puis à faire un filtrage bidimensionnel qui agit à la fois en temps et sur l'échelle. Ainsi, la modélisation de la réponse à un signal sonore peut être schématisé comme suit:

$$x(t) \xrightarrow{cochlée} x(t) * \psi_s(t) \xrightarrow{cortex(A1)} \rho[\rho[x(t) * \psi_s(t)] * \psi_\alpha(t, s)]$$

Dans la partie A2 du cortex auditif cela devient beaucoup plus compliqué avec des sortes d'analyse de patterns. Ce que l'on constate, c'est que plus on s'enfonce dans le cortex profond, plus les coefficients évoluent lentement dans le temps, comme une invariance dans le temps sur des plages de la seconde ou plus.

En se référant au séminaire de Geoffroy Peters (12 février 2020) toutes les représentations qu'il a présentées sont des transformées en ondelettes, qu'il a nommé par le vocable *"the Q constant transform"*. La Q-value est la largeur de l'ondelette en représentation $\log \omega$, laquelle est constante quand on change l'échelle s, comme illustré sur la figure 39.

Pour se donner une intuition de la largeur Q dans le cas de l'oreille, envisageons la notion d'octave. Le découpage de la bande de fréquences en intervalles de largeur *relative* constante, ou $\Delta\omega/\omega = \text{Cte}$ définit d'une manière général une octave [50]. La largeur de la

50. NDJE: Typiquement la bande d'une octave est définie par une fréquence centrale ω_c, et une fréquence ω appartient à l'octave si $\log \omega = \log w_c \pm \log \sqrt{2}$. Donc, $\Delta \log f = \Delta f/f = \log 2$.

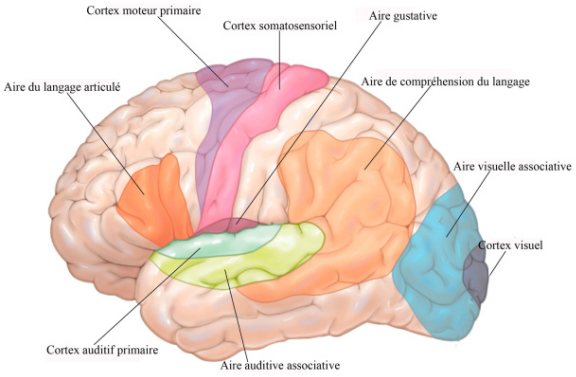

FIGURE 38 – Zones spécialisées du cortex cérébral (gauche).

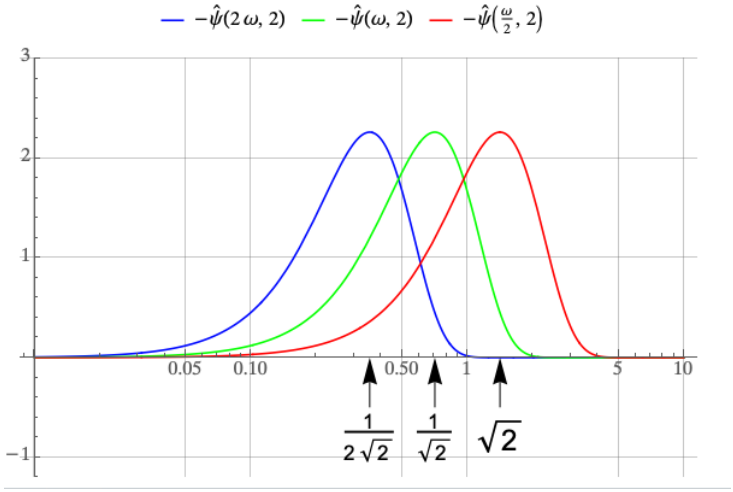

FIGURE 39 – Illustration de la constance de la largeur de l'ondelette en fréquence, dans une représentation logarithmique. (Voir Fig. 33).

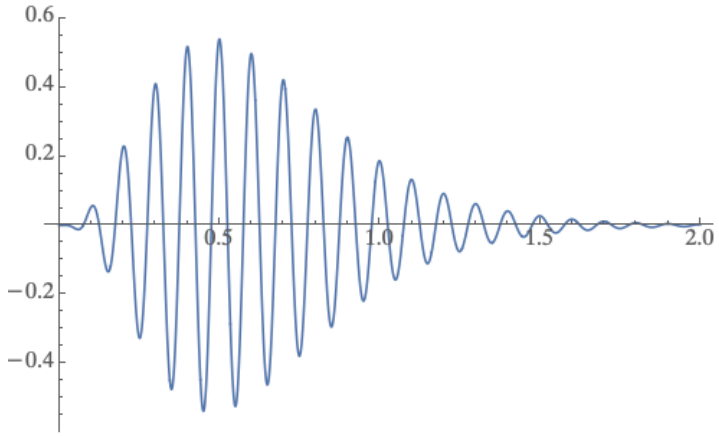

FIGURE 40 – Exemple de "gammatone" d'expression générique $at^{n-1}e^{-2\pi bt}\cos(2\pi ft)$.

bande d'une octave est donc égale à la distance à la fréquence 0. La gamme dite à "tempérament égal" divise 1 octave en 12 (demi-tons) intervalles en progression géométrique que l'oreille humaine distingue. Donc, on a une meilleure résolution que l'octave. En fait quand on mesure la largeur des filtres cochléaires, on obtient $\sim 1/16$ d'octave. Les ondelettes sont des filtres "gammatone" (Fig 40) qui ont une forme asymétrique en temps pour capturer la non-réversibilité du signal. Cependant, la largeur en fréquence du gammatone est bien plus petite que sa distance à 0 [51].

Ces ondelettes en audio sont différentes de celles rencontrées en image mais les propriétés mathématiques restent essentiellement les mêmes. Le choix de l'ondelette est dictée par la meilleure parcimonie de la représentation.

7.4 Descripteurs MFC

En analyse audio, les spécialistes du traitement du signal ont définit dans les années 1970-80 des descripteurs MFC (Mel-Frequency Cepstrum). Ces MFC ont été utilisés systématiquement dans tous les algorithmes jusqu'en 2010. D'un certain point de vue, les MFC sont issus d'une technologie qui n'a pas évoluée pendant 30-40 ans. Celle-ci a vu

51. NDJE: cette remarque fait référence à la façon dont on peut construire une ondelette complexe analytique à partir d'une gaussienne de largeur η (en Fourier), il suffit de la translatée son spectre de la même quantité vers les hautes fréquences.

surgir les CNN qui à partir de 2014-15 sont devenus les plus performants. Ceci dit, ces MFC sont bien compris analytiquement et donnent une bonne base pour comprendre les premières couches des CNN. Passons en revu comment on obtient ces MFC:

1. On commence par obtenir $\mathrm{Mel}(x)$ en calculant **les modules de coefficients d'ondelettes** à une échelle $s = a^j$, c'est-à-dire $|x * \psi_{a^j}|$ [52]. Le coefficient a doit pourvoir faire en sorte que les filtres couvrent toute la bande des fréquences. Notons que $\log s = j \log a$, ce qui revient à dire que l'on échantillonne uniformément l'axe des fréquences, car pour rappel la fréquence centrale de la boite d'introspection est $\omega_c = \xi/s$, donc $\log \omega_c = \log \xi - j \log a$ avec, $\log \xi$ et $\log a$ deux constantes. Ensuite, **on effectue une convolution** qui délocalise sur une échelle de temps 2^J via $|x * \psi_{a^j}| * \phi_J(t)$ avec $\phi(t)$ une fenêtre de largeur ~ 1 que l'on dilate pour ajuster sa durée, cf. $\phi_J(t) = 2^{-J}\phi(2^{-J}t)$ de largeur $\sim 2^J$, typiquement de l'ordre de 25ms à 50ms. Ces descripteurs vont avoir un impact très important concernant **l'invariance locale par translation ϕ_J et stabilité par déformation avec les ondelettes** ψ_{a^j}.
2. une fois obtenus ces descripteurs Mel qui filtrent le signal dans des bandes de fréquences de largeur constante en échelle logarithmique, pour construire des représentations qui sont insensibles aux variations du signal non-informatives, on va construire des **invariants**.

   Le premier invariant auquel on pense, est donné par le fait que l'on veut s'affranchir d'un **facteur multiplicatif du signal**. Or, si $x(t) \to \alpha x(t)$ alors $\mathrm{Mel}(x) \to \alpha \mathrm{Mel}(x)$. Pour isoler le facteur inconnu α, on peut penser à prendre le logarithme, qui donne:

$$\log(\alpha \mathrm{Mel}(x)) = \log \alpha + \log(\mathrm{Mel}(x)) \tag{105}$$

   Ensuite, on veut s'affranchir d'une **modulation d'amplitude**, qui est une extension du cas précédent, en laissant α dépendre du temps, mais à peu prés constant sur des intervalles de durée 2^J. Donc si $\alpha(t)$ a des **variations lentes** par rapport à celles de ψ_{a^j} et ϕ_J, alors on peut le sortir des convolutions et donc

$$\log(\alpha \mathrm{Mel}(x)) = \log \alpha(t) + \log(\mathrm{Mel}(x)) = \log \alpha(t) + \log[|x * \psi_{a^j}| * \phi_J(t)] \tag{106}$$

52. On peut prendre le module carré aussi mais comme on le verra on prend par la suite le log du Mel donc la puissance n'a pas d'importance.

Il y a **deux composantes qui dépendent du temps** qu'il faut séparer car elles contiennent des informations de différentes natures: la première ($\log \alpha(t)$) ne dépend que du temps, par contre la seconde dépend à J fixé (cf. la résolution temporelle) de j et de t. On effectue alors **une transformation le long de l'axe de** $\log \omega$ **qui sépare la partie constante (cf. indépendante de** j**) et la partie qui varie**. Cette transformation ne peut être une DFT [53] car le "signal" le long de l'axe $\log \omega$ n'est pas périodique, ce qui aurait des effets de bord nuisibles. Par contre, **on effectue une DCT** (Discrete Cosine Transform) [54]:

$$DCT \rightarrow \left\langle z(j), c_k(j) = \cos\left(\frac{i2\pi k}{K}(j+1/2)\right)\right\rangle \tag{107}$$

avec $z(j) = \log|x * \psi_{a^j}| * \phi_J(t)$ à t fixé. On peut démontrer que les $\{c_k(j)\}_{0\leq k<K}$ définissent une base orthogonale. Donc, on calcule

$$a_k = \sum_j z(j)c_k(j) \tag{108}$$

c'est-à-dire que l'on extrait une information sur **les variations du signal le long de l'axe des fréquences**.

Le résultat que l'on appelle un MFC, c'est une collection de MFCCs (Mel-frequency cepstral coefficients) qui dépendent d'un temps t et d'une fréquence de fréquence k (nb. j relie l'échelle logarithmique de fréquence de l'ondelette ψ_{a^j}, et J l'échelle de la fonction de lissage $\phi_J(t)$)

$$\boxed{MFCC(t,k) = \langle \log \mathrm{Mel}_x(t,j,J), c_k(j)\rangle = DCT_j(\log \mathrm{Mel}_x)(t,k)} \tag{109}$$

Le paramètre j est un paramètre de fréquence, k un paramètre de fréquence de fréquence, et t est échantillonné avec l'échelle 2^J (cf. $t_n = n2^J$). Un schéma de ces différentes étapes est présenté sur la figure 41. Finalement pour résumer: **tous les 25ms, on dispose d'un vecteur qui selon** k **donne la variation du module de la transformée en Ondelettes lissée sur 25ms**.

53. NDJE: DFT pour Discrete Fourier Transform, pour distinguer de FFT pour Fast Fourier Transform, bien que dans la pratique on utilise le vocable FFT pour désigner la DFT.

54. NDJE: il y a différentes variantes de DCT, donc il faut se pencher sur les documentations des librairies utilisées.

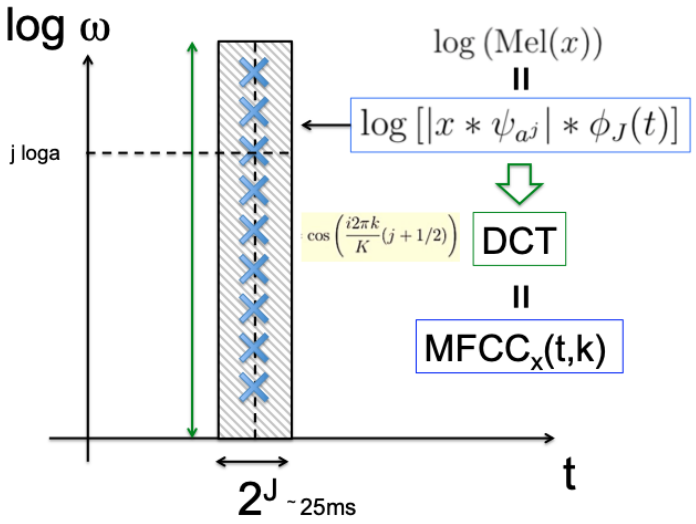

FIGURE 41 – Schématisation des différentes étapes pour obtenir les MFCC du signal x. Elles font appel à la transformée en ondelettes ψ_{a^j} lissée dans le domaine temporel par $\phi_J(t)$, puis à une DCT sur le long de l'axe des fréquences via les cosinus $c_k(j)$ (Eq. 109).

La valeur pour $k = 0$, le MFCC correspondant est dépendant des modulations de fréquences lentes, on les isole donc dans 1 seul coefficient. Tandis que pour tous les autres coefficients ($k > 0$) on voit apparaître la signature des harmoniques. Par la suite, ces coefficients $MFCC(n2^J, k)$ sont soit envoyés directement dans un classificateur par exemple pour faire de la reconnaissance de genres musicaux, soit on construit des algorithmes de chaines de Markov pour faire de la reconnaissance de la parole pour connecter les $MFCC(n2^J, k)$ selon n à l'aide de modèles Gaussiens (GMM) qui par la suite sont envoyés dans des classificateurs. Ces développements étaient l'état de l'art jusque dans les années 2010.

7.5 Inversion et stabilité de la transformée en Ondelettes

Nous allons aborder, pourquoi il a été fondamental d'utiliser les ondelettes, et discuter les propriétés de stabilité quand on fait de la reconnaissance. Dans un premier temps, nous allons démontrer les propriétés d'inversion des ondelettes, puis dans un second temps nous étudierons leur stabilité par déformation.

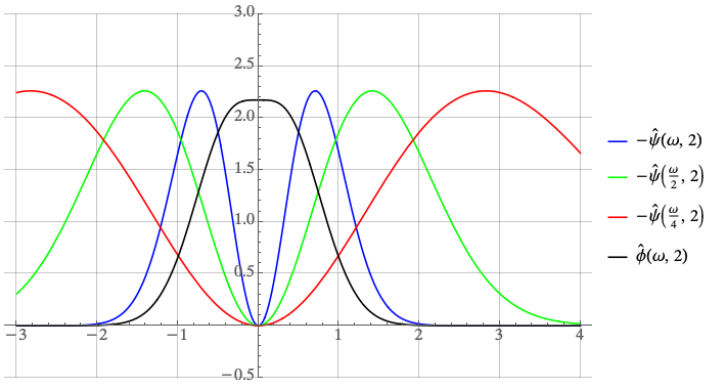

FIGURE 42 – Illustration de la représentation des filtres d'ondelettes $\hat{\psi}$ (Eq. 103) et du filtre passe-bas $\hat{\phi}$.

Pour le calcul d'un MFC, on utilise la transformée en ondelettes ψ_{a^j} avec une échelle $s = a^j$. Mais, on veut aussi capturer les très basses fréquences [55]. Pour ce faire, on va utiliser un filtre spécial $\phi(t)$ que l'on dilate dans le temps $\phi_J(t) = 2^{-J}\phi(2^{-J}t)$ (rappel: $\hat{\phi}_J(\omega) = \hat{\phi}(2^J\omega)$). Donc, on dispose d'une collection de filtrages de x à basses et hautes fréquences, selon

$$\boxed{Wx \equiv \Big((x * \psi_{a^j})(t), (x * \phi_J)(t)\Big)_{a^{-j} \geq 2^{-J}}} \tag{110}$$

et toutes les échelles a^j sont plus petites que 2^J, ou $j \leq J(\log 2/\log a) = \alpha J$. Sur la figure 42, on voit que si $J = 0$ (cf. $\hat{\phi}_0 = \hat{\phi}$), on retient tous filtres d'ondelettes pour lesquels $2^j : j = 0, -1, -2, \ldots$ (avec $a = 2$).

Les deux premières questions qui viennent à l'esprit sont de savoir: 1) si à partir des W_x on peut synthétiser $x(t)$ et 2) est-ce que cette transformation est stable, c'est-à-dire que $\|Wx\| \sim \|x\|$? On va se placer dans le cas où $\psi(t)$ est une **ondelette réelle**, c'est-à-dire pour laquelle $\hat{\psi}(-\omega) = \hat{\psi}^*(\omega)$. Par la suite nous verrons le cas d'une **ondelette complexe et analytique**, c'est-à-dire que $\hat{\psi}(\omega)$ a une support borné sur $\mathbb{R}^+$ [56].

55. NDJE: ici la fonction ϕ a l'unique fonction de couvrir le spectre à basse fréquence. Dans le cours de 2018, et dans son livre, S. Mallat introduit la fonction de scaling qui impose une forme de filtre particulier car il est relié à celui de ψ (et vice-versa). Dans les figures pour illustration j'ai choisi la fonction ϕ associée au chapeau mexicain.

56. NDJE: S. Mallat n'a pas eu le temps de traiter ce dernier cas de figure durant son cours oral.

Théorème 7. *Dans le cas d'une ondelette* ***réelle****, on a besoin que l'ensemble de filtres* $\left\{\hat{\psi}(a^j\omega)\right\}_{j\leq\alpha J}$ *et* $\hat{\phi}(2^J\omega)$ *satisfassent la relation:*

$$\exists c<1 \text{ tq. } \forall\omega,\ 1-c\leq S(\omega)=\sum_{j\leq\alpha J}|\hat{\psi}(a^j\omega)|^2+|\hat{\phi}(2^J\omega)|^2\leq 1 \tag{111}$$

alors on définit deux filtres $\widehat{\overline{\psi}}(\omega)$ *et* $\widehat{\overline{\phi}}(\omega)$ *selon*

$$\widehat{\overline{\psi}_{a^j}}(\omega)\equiv\frac{\hat{\psi}^*(a^j\omega)}{S(\omega)},\quad \widehat{\overline{\phi}_J}(\omega)\equiv\frac{\hat{\phi}^*(2^J\omega)}{S(\omega)} \tag{112}$$

tels que

$$\boxed{x(t)=\underbrace{\textstyle\sum_j(x*\psi_{a^j}*\overline{\psi}_{a^j})(t)}_{\text{hautes fréquences}}+\underbrace{(x*\phi_J*\overline{\phi}_J)(t)}_{\text{basses fréquences}}} \tag{113}$$

et

$$\boxed{(1-c)\|x\|^2\leq\|Wx\|^2=\textstyle\sum_j\|x*\psi_{a^j}\|^2+\|x*\phi_J\|^2\leq\|x\|^2} \tag{114}$$

La contrainte (Eq. 111) stipule en fait que le spectre couvert par l'ensemble des filtres n'a pas de trous. Un exemple avec les filtres en chapeau mexicain et le filtre passe-bas associé est donné sur la figure 43.

Démonstration 7. Naturellement avec des sommes de convolution, la démonstration se fait dans l'espace de Fourier. Notons $b(t)$ le membre de droite (r.h.s) de l'équation 113 et prenons sa transformée de Fourier:

$$\begin{aligned}\hat{b}(\omega)&=\textstyle\sum_j\hat{x}(\omega)\widehat{\psi_{a^j}}(\omega)\widehat{\overline{\psi}_{a^j}}(\omega)+\hat{x}(\omega)\widehat{\phi_J}(\omega)\widehat{\overline{\phi}_J}(\omega)\\&=\hat{x}(\omega)\left\{\textstyle\sum_j\dfrac{|\hat{\psi}(a^j\omega)|^2}{S(\omega)}+\dfrac{|\hat{\phi}(2^J\omega)|^2}{S(\omega)}\right\}\\&=\frac{\hat{x}(\omega)}{S(\omega)}\left\{\textstyle\sum_j|\hat{\psi}(a^j\omega)|^2+|\hat{\phi}(2^J\omega)|^2\right\}=\hat{x}(\omega)\end{aligned} \tag{115}$$

Pour la seconde relation (dite de *la conservation de l'énergie*), on procède en utilisant

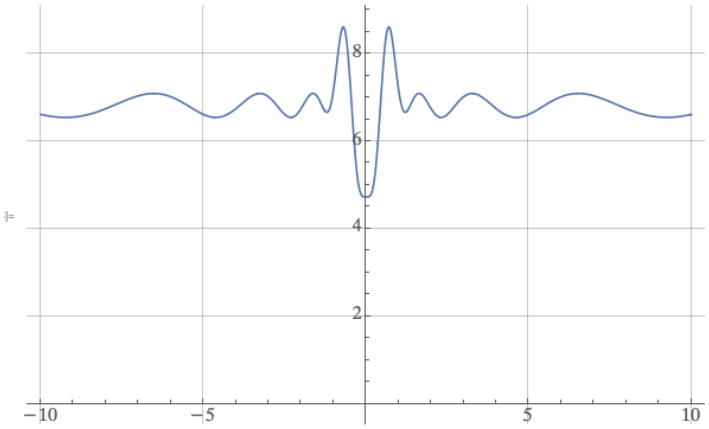

FIGURE 43 – Exemple de $S(\omega)$ (Eq. 111) dans le cas des $\hat{\psi}$ (Eq. 103) et du filtre passe-bas $\hat{\phi}$. En fait, l'important est que $S(\omega)$ n'ait pas de trous, et la borne supérieure de 1 est conventionnelle.

Parseval. Or,

$$\begin{aligned}\|Wx\|^2 &= \frac{1}{2\pi}\left\{\sum_j \int |\hat{x}(\omega)|^2 |\hat{\psi}(a^j\omega)|^2 d\omega + \int |\hat{x}(\omega)|^2 |\hat{\phi}(2^J\omega)|^2 d\omega\right\} \\ &= \frac{1}{2\pi}\int |\hat{x}(\omega)|^2 \left\{\sum_j |\hat{\psi}(a^j\omega)|^2 + |\hat{\phi}(2^J\omega)|^2\right\} d\omega \\ &= \frac{1}{2\pi}\int |\hat{x}(\omega)|^2 S(\omega) d\omega \end{aligned} \tag{116}$$

et en utilisant la contrainte Eq. 111 sur $S(\omega)$, on obtient le résultat. ■

Donc, ce théorème dit qu'une fois que les filtres couvrent toutes les fréquences, la décomposition est bien complète et stable.

Enfin, le facteur a pour le traitement d'image on verra qu'il est égal à 2, dans l'audio $a = 2^{1/Q}$ avec Q la fameuse largeur du filtre qui typiquement est égal à $8, 16, 32$. C'est-à-dire que l'on met Q fréquences (demi-tons) par octave.

7.5.1 Le cas des ondelettes analytiques

NDJE: c'est un cas que S. Mallat n'a pu traiter en cours faute de temps. Voici donc un complément qu'il a bien voulu valider.
Par définition, une ondelette ψ dont le spectre de Fourier est non nul uniquement pour $\omega \leq 0$ soit

$$\hat{\psi}(\omega) = 0 \quad \text{si} \quad \omega < 0 \tag{117}$$

est une fonction *complexe analytique*. La condition Eq. 111 n'est valable que pour $\omega \geq 0$. Si on note $b(t)$ la partie droite de Eq. 113, alors $\hat{x}(\omega) = \hat{b}(\omega)$ également pour $\omega \geq 0$, et $\hat{b}(\omega) = 0$ pour $\omega < 0$. Pour **un signal** $x(t)$ **réel** pour les $\omega < 0$, on sait que $\hat{x}(-\omega) = \hat{x}^*(\omega)$ donc $\hat{x}(\omega) = \hat{b}^*(-\omega)$ pour $\omega < 0$. Donc, pour reconstruire $x(t)$ par Fourier Inverse, on a

$$\begin{aligned} x(t) &= \frac{1}{2\pi}\left\{\int_0^{+\infty} \hat{b}(\omega)e^{i\omega t}d\omega + \int_{-\infty}^{0} \hat{b}^*(-\omega)e^{i\omega t}d\omega\right\} \\ &= \frac{1}{2\pi}\left\{\int_{-\infty}^{+\infty} \hat{b}(\omega)e^{i\omega t}d\omega + \int_{-\infty}^{+\infty} \hat{b}^*(\omega)e^{-i\omega t}d\omega\right\} \\ &= b(t) + b^*(t) = 2\,\mathrm{Re}[b(t)] \end{aligned} \tag{118}$$

Donc, la synthèse du signal réel, dans le cas d'une ondelette analytique, se fait par la relation

$$\boxed{x(t) = 2\,\mathrm{Re}\left[\sum_j (x * \psi_{a^j} * \overline{\psi}_{a^j})(t) + (x * \phi_J * \overline{\phi}_J)(t)\right]} \tag{119}$$

Pour ce qui est de la conservation de l'énergie, par le même calcul Eq. 116, on obtient une relation valable uniquement pour les fréquences positives, à savoir:

$$\|Wx\|^2 = \frac{1}{2\pi}\int_0^{+\infty} |\hat{x}(\omega)|^2 S(\omega)d\omega \leq \frac{1}{2\pi}\int_0^{+\infty} |\hat{x}(\omega)|^2 d\omega \tag{120}$$

Or, pour un **signal réel**

$$\|x\|^2 = \frac{2}{2\pi}\int_0^{+\infty} |\hat{x}(\omega)|^2 d\omega \tag{121}$$

donc, on obtient la double contrainte dans le cas d'une ondelette analytique et un signal réel:

$$\boxed{(1-c)\frac{\|x\|^2}{2} \leq \|Wx\|^2 \leq \frac{\|x\|^2}{2}} \tag{122}$$

7.6 La représentation $\Phi(x)$

Gardons à l'esprit que l'on veut construire une représentation $\Phi(x)$ qui élimine la malédiction de la dimension en utilisant les $3S$ (séparabilité, symétrie, sparsité). Donc, à partir des MFC que doit-on faire? Pour se concentrer sur l'essentiel, on va oublier le "log" ainsi que la DCT, et on va montrer que si:

$$\boxed{\Phi(x) \equiv \{\rho(x * \psi_{a^j}) * \phi_J(t)\}_j} \tag{123}$$

où ρ est soit la valeur absolue soit un ReLU (ou tout autre opérateur contractant), alors on a une représentation

1. qui n'amplifie pas le bruit (cf. *propriété de contractance*);
2. et stable aux déformations.

Théorème 8. *On reprend les hypothèses du théorème 7, alors il vient que* Φ *est* ***contractante****, à savoir que*

$$\forall x, x' \quad \|\Phi(x) - \Phi(x')\| \leq \|x - x'\| \tag{124}$$

et comme $\Phi(0) = 0$*, alors il vient comme conséquence que*

$$\|\Phi(x)\| \leq \|x\| \tag{125}$$

(nb. sur la première inégalité, on peut mettre une constante).

En classification, c'est la première propriété qu'il faut assurer car en grande dimension les points sont très loin les uns des autres, donc on recherche à obtenir une représentation qui les rapproche au maximum, surtout en rapprochant des points appartenant à la même classe.

Démonstration 8. La démonstration commence par faire remarquer que $\Phi(x)$ mobilise 3 opérations: 1) la transformation en ondelettes qui est linéaire, 2) la prise du module qui se traduit par prendre une fonction $\rho(z) = |z|$ (module de "z" si a est complexe), notons que cela pourrait être un rectificateur ReLU car c'est aussi un opérateur contractant, 3) le lissage par Φ_J est une moyenne de type $A_J(z) = z * \Phi_J$. Donc, on peut écrire $\Phi(x)$ selon

$$\Phi(x) = A_J[\rho(Wx)] \tag{126}$$

Ce que l'on sait d'après le théorème 7 c'est que W est un opérateur linéaire pour lequel

$$\|Wx\| \leq \|x\| \Rightarrow \|W(x-x')\| = \|Wx - Wx'\| \leq \|x - x'\|$$

Donc W est un opérateur contractant [57]. Ensuite, pour l'opérateur ρ (module ou ReLU), on sait que la valeur absolue satisfait l'inégalité triangulaire renversée (c'est vrai aussi pour le ReLU)

$$|\rho(z) - \rho(z')| \leq |z - z'|$$

donc ρ est un opérateur contractant. Il suffit maintenant de se concentrer sur A_J.

Lemme: Soit l'opérateur A_J linéaire définit par $A_J z = z * \Phi_J$, pour montrer qu'il est contractant, il suffit de montrer que

$$\|A_J z\| \leq \|z\|$$

Or, A_J est un opérateur de convolution, donc on passe dans le domaine de Fourier, et en utilisant la seconde inégalité Eq. 114 du théorème 7, alors comme $|z * \Phi_J||^2 \leq ||z||^2$. Donc, A_J est bien un opérateur contractant.

Finalement, l'enchainement d'opérateurs $A_J \rho W$ est contractant, donc le théorème est démontré. ■

Notons que **la contraction** ne va pas être forte au niveau de W car il conserve la norme, mais elle **va être forte au niveau de l'opérateur ρ (ex. le ReLU) et au niveau du moyennage (*pooling*, en réseau de neurones on utilise le "max pooling" ou "l'average pooling")**. Ceci dit dans cette opération de contraction, on peut très bien rapprocher des x qui ne sont pas de la même classe donc faudra des mécanismes pour empêcher cela. Ceci dit, on peut cascader ces opérations pour contracter un maximum. Abordons, le résultat de la stabilité par déformation qui est un peu nouveau. En effet, en analyse harmonique on s'est concentré sur des opérateurs linéaires, et donc des résultats sur des opérateurs non-linéaires sont relancés pour comprendre les MFCC et les CNN.

57. Attention, pour un opérateur non-linéaire, on peut avoir la propriété de gauche sur la norme de $O_{NL}(x)$ sans avoir la propriété de "contractance"

Théorème 9. *Il s'agit de démontrer la continuité Lipschitzienne par petite déformation, c'est-à-dire que*[58]

$$g.x(u) = x(u - \tau(u)), \text{ avec } ||\nabla\tau||_\infty < 1/2$$

Donc, si on se restreint à des signaux $x \in L^2(\Omega)$ sur un support compact Ω[59] *qui peuvent être très irréguliers, alors:*

$$\exists C > 0 \text{ tq. } \forall \tau \in C^2 \ ||\nabla\tau||_\infty < 1/2, \forall x \in L^2(\Omega),$$
$$\|\Phi(x) - \Phi(g.x)\| \leq C\|x\| \left(\underbrace{||\nabla\tau||_\infty + \|H\tau\|_\infty}_{|g|_G} + \underbrace{\|\tau\|_\infty 2^{-J}}_{\text{translation}} \right) \quad (127)$$

avec $\|H\tau\|_\infty$ la norme infinie du Hessien (cf. les dérivées 2nd), et $\|\tau\|_\infty = \sup_u |\tau(u)|$. Notons que la norme de $\Phi(x)$, comme c'est un vecteur, on a

$$\|\Phi(x)\|^2 = \sum\nolimits_j \|\rho(x * \psi_{a^j}) * \phi_J(t)\|^2$$

Ce que le théorème nous dit est que **lorsque $J \to \infty$ alors la composante de translation va disparaître et la borne ne va dépendre que de la taille de la déformation** (à rapprocher du raisonnement Eq. 43). C'est un résultat fort car il porte sur toute la représentation. Si la déformation est petite alors la représentation linéarise ces déformations (stabilité), donc on pourra sélectionner et apprendre des invariants sur les petites déformations en apprenant des opérateurs linéaires, ce que peut faire un classificateur linéaire appliqué sur $\Phi(x)$. Le problème est ici de linéariser approximativement les sources de variabilités que l'on veut éliminer. Par contre, si la déformation est trop grande alors cette linéarisation n'est plus valable.

NDJE: la démonstration est reportée à la séance suivante Sec. 8.5.

58. La condition porte sur la norme de la matrice jacobienne en dimension 2, ... en 1d il s'agit de simplement que $|\tau'(u)| < 1$. Ce faisant même si on peut aller jusqu'à 1, pour enlever toute instabilité autour de 1, on fixe la borne à 1/2.

59. On peut étendre à tout $\mathbb{R}$ mais çà entraine des complications techniques

8. Séance du 11 Mars

Lors de cette séance, nous allons dans un premier temps reprendre les descripteurs calculés à partir des transformations en ondelettes qui sont covariants, invariants par translation multi-échelles. Nous allons voir comment ils se comportent dans le contexte du traitement d'images et nous verrons le lien avec la neurophysiologie. Dans un second temps, nous ferons une démonstration de la covariance et stabilité par difféomorphisme de ces descripteurs qui permet de les linéariser pour la classification (thème important). La troisième et dernière partie sera consacrée au lien avec les réseaux de neurones, en montrant comment ces descripteurs peuvent être implémentés en utilisant des cascades de filtres (convolution/pooling et ReLU). Ainsi, cela sera une première façon d'aborder l'architecture des réseaux CNN, avec une architecture ultra-simplifiée. La semaine suivante (sic) sera/serait[60] *dédiée aux applications de ce genre d'architecture.*

8.1 Rappel des MFC en audio

On prend une familles d'ondelettes ψ_j définies par

$$\psi_j(u) = \frac{1}{a^j}\psi\left(\frac{u}{a^j}\right) \Rightarrow \hat{\psi}_j(\omega) = \hat{\psi}(a^j\omega) \tag{128}$$

L'ondelette ψ est choisie pour réaliser un filtre passe-bande lequel est dilaté par le facteur d'échelle a^j (voir la figure 33 par exemple). Dans l'audio le facteur a est plus petit que 2, à savoir $a = 2^{1/Q}$ où Q est la largeur du filtre en échelle logarithmique (Fig. 39). Pour couvrir une octave (cf. un facteur 2), il faut Q ondelettes. Donc, Q donne la précision fréquentielle, plus Q est grand, meilleure est la résolution. Typiquement, $Q \sim 16$, c'est-à-dire un peu plus grand que le nombre de demi-tons en musique, c'est aussi ce que l'on retrouve comme précision des filtres de la cochlée.

La transformée en ondelettes est définie comme une collection de convolutions à la fois avec la famille d'ondelettes ψ_j pour couvrir les hautes fréquences, et un filtre complémen-

60. NDJE: au moment où S. Mallat fait son cours, il ne peut anticiper les mesures draconiennes qui seront mises en place pour minimiser au maximum les impacts de la pandémie due au Covid-19.

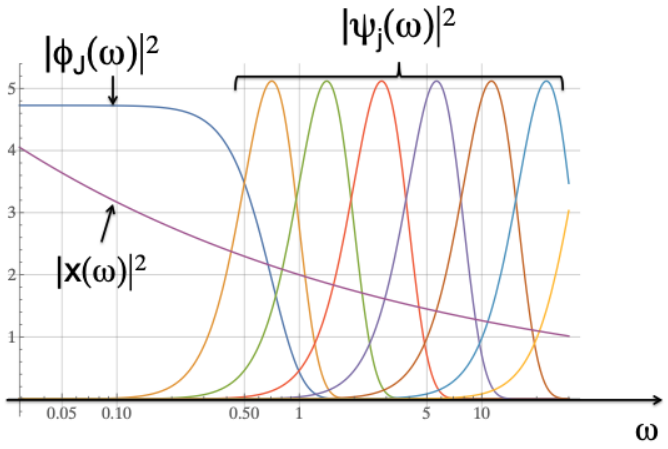

FIGURE 44 – Illustration de la représentation des filtres d'ondelettes $\hat{\psi}$ (Eq. 103) et du filtre passe-bas $\hat{\phi}$, en échelle logarithmique des fréquences, qui permet de cerner les différentes parties du signal x qui sont introspectées par les différents filtres.

taire $\phi_J = 2^{-J}\phi(2^{-J}u)$ qui prend en charge les basses fréquences (voir Sec. 7.5)

$$Wx = \left(\{x * \psi_j\}_j, x * \phi_J\right) \Rightarrow \widehat{Wx}(\omega) = \left(\left\{\hat{x}(\omega)\hat{\psi}(a^j\omega)\right\}_j, \hat{x}(\omega)\hat{\phi}(2^J\omega)\right) \tag{129}$$

Cela revient à analyser le signal dans toutes les bandes de fréquences comme illustré sur la figure 44. Rappelons que ψ est d'integrale nulle donc par nature elle oscille parfois beaucoup, ce qui n'est pas forcement le cas de ϕ [61].

On a vu que ce genre de représentation est **complète** et **stable** à condition que l'on couvre tout **l'axe des fréquences sans trou**. Cette condition se traduit par la relation

$$0 < 1 - c \leq S(\omega) \equiv \sum_j |\hat{\psi}(a^j\omega)|^2 + \hat{\phi}(2^J\omega) \leq 1 \tag{130}$$

avec la partie importante est que la borne inférieure et **strictement supérieure** à 0 (la borne sup. est normalisée à 1 par convention). Ceci a deux conséquences fondamentales:

— **complétude**: on peut définir une ondelette de reconstruction $\overline{\psi}$ pour les hautes fré-

61. NDJE: dans la littérature, il y a des collections d'ondelettes ψ et de leurs *scaling functions* associées avec des caractéristiques très différentes de celles présentées par S. Mallat, donc il faut se placer dans l'optique de ce cours. Dans le cas de ψ en chapeau mexicain, la fonction ϕ a également une forme identique mais l'undershoot est bien moins prononcé.

quences et son pendant pour les basses fréquences $\overline{\phi}$ définies par leurs transformées de Fourier selon

$$\widehat{\overline{\psi}_{a^j}}(\omega) \equiv \frac{\hat{\psi}^*(a^j\omega)}{S(\omega)}, \quad \widehat{\overline{\phi}}_J(\omega) \equiv \frac{\hat{\phi}^*(2^J\omega)}{S(\omega)} \tag{131}$$

ainsi

$$\boxed{x(t) = \underbrace{\sum\nolimits_j (x * \psi_{a^j} * \overline{\psi}_{a^j})(t)}_{\text{hautes fréquences}} + \underbrace{(x * \phi_J * \overline{\phi}_J)(t)}_{\text{basses fréquences}}} \tag{132}$$

— **opérateur contractant**: la transformée en ondelettes satisfait la double inégalité

$$\boxed{(1-c)\|x\|^2 \leq \|Wx\|^2 = \sum\nolimits_j \|x * \psi_j\|^2 + \|x * \phi_J\|^2 \leq \|x\|^2} \tag{133}$$

A partir de la transformée en ondelettes Wx, on a vu que l'on définit des descripteurs de Mel fréquence (MFC) selon (nb. on omet la partie basse fréquence pour alléger les notations):

$$\boxed{\Phi(x) = \{\rho(x * \psi_{a^j}) * \phi_J(t)\}_j} \tag{134}$$

avec ρ **une non linéarité** (module, ReLU). L'idée pour étudier ces descripteurs qui ont été supplantés dans les années 2010, c'est que l'on a **tous les ingrédients de base des réseaux de neurones**:

1. on choisit l'ondelette ψ pour avoir une **représentation parcimonieuse** du signal lors de l'opération $x * \psi_{a^j}$ (Wx), qui de plus est une opération **covariante**[62] par translation;
2. on effectue une **rectification non-linéaire** par ρ qui est aussi un opérateur **covariant**;
3. la troisième opération est donnée par la moyenne (**pooling**) ϕ_J (A_J) qui est une opération **invariante**

qui permettent d'écrire la cascade d'opérateurs pour calculer $\Phi(x)$ selon

$$\boxed{\Phi(x) = (A_J \rho W)(x)} \tag{135}$$

La non-linéarité est indispensable! Imaginons que l'on supprime l'action de ρ, alors on

62. NDJE: rappel covariant et équivariant Sec. 5.7.

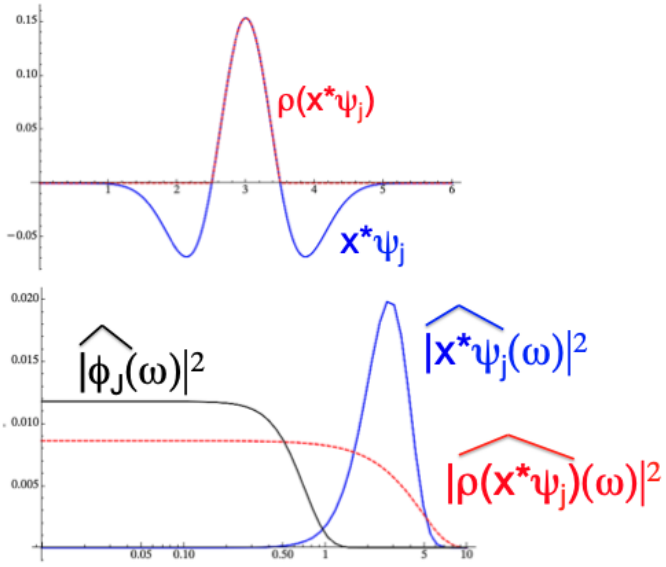

FIGURE 45 – Illustration de l'effet d'un ReLU sur la convolution $x * \psi_{a^j}$. En haut, le résultat dans l'espace réel, et en bas ce qui se passe dans l'espace de Fourier. Avec le ReLU, la composante basse fréquence produit avec le filtre passe-bas ϕ_J donne une contribution non nulle (ici pour faire simple on a pris $x(u) = \delta(u - u_0)$ donc la convolution redonne l'ondelette translatée.

moyenne les coefficients d'ondelette ce qui donne 0. En effet, en Fourier

$$x * \psi_{a^j} * \phi_J(t) \xrightarrow{T.F} \hat{x}(\omega)\hat{\psi}(a^j\omega)\hat{\phi}(2^J\omega) \tag{136}$$

or $\hat{\phi}(2^J\omega)$ est un filtre basses-fréquences alors que les $\hat{\psi}(a^j\omega)$ sont des filtres passe-bandes à hautes-fréquences, donc leur produit est quasi nul. Par contre, quand on applique le ReLU (par ex.), celui-ci ne garde que la partie positive qui va faire en sorte qu'il y aura un recouvrement non nul (Fig. 45). C'est la non-linéarité qui permet d'avoir **des nouveaux invariants** qui ne sont pas uniquement la moyenne (rappel: le seul opérateur linéaire invariant par l'action d'un groupe, c'est la moyenne). Ici la moyenne est appliquée après la non-linéarité ce qui donne de nouvelles propriétés:

— $\Phi(x)$ est **contractant** (voir Th. 8). Ceci vient du fait que chacun des trois opérateurs W, ρ et A_J est contractant.

— $\Phi(x)$ est **stable par déformation** ce que nous allons voir dans une section ultérieure.

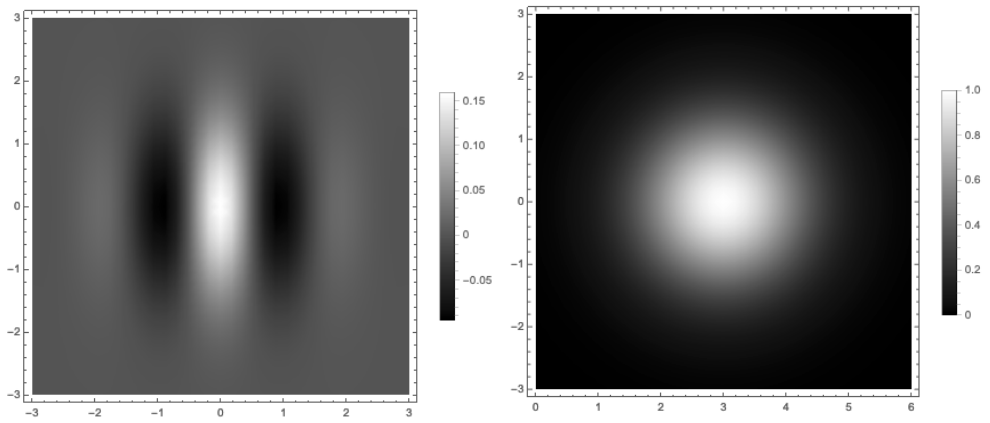

FIGURE 46 – Exemple d'ondelette bidimensionnelle $\mathcal{N}_{(0,1)}(x,y)e^{i3x}$: à gauche la partie réelle, à droite la transformée de Fourier (notez que l'on obtient un gaussienne décalée selon l'axe horizontal de 3 unités.

8.2 Les descripteurs sur des images

Avant d'aborder la stabilité de $\Phi(x)$, nous allons voir comment ces descripteurs se déclinent pour des images. Dans ce cas de figure, les ondelettes sont des objets bidimensionnels. Tout d'abord, comme pour l'audio, on définit ψ selon

$$\psi(u) = g(u)e^{i\xi.u} \xrightarrow{T.F} \hat{\psi}(\omega) = \hat{g}(\omega - \xi) \tag{137}$$

avec g une gaussienne (ou autre fenêtre régulière), et ξ une "fréquence" fixe. Voir un exemple sur la figure 46. Donc, on veut un opérateur W (transformée en ondelettes) qui **sparsifie le signal**, c'est-à-dire qui puisse capturer uniquement que les structures pertinentes (contours avec leurs orientations...), et on veut également qu'il soit **complet**, donc il faut **pouvoir couvrir tout le domaine de "fréquences"**. Pour ce faire, on va introduire une **rotation des ondelettes**:

$$\psi_\theta(u) \equiv \psi(r_{-\theta}.u) \tag{138}$$

Ensuite, selon le principe des ondelettes 1D, on va dilater/compresser l'échelle pour avoir une palette complète de filtres: des petits pour les petites structures, et des grands pour

$$\psi_{\theta,j}(u) \equiv 2^{-2j}\psi(2^{-j}(r_{-\theta}.u))$$

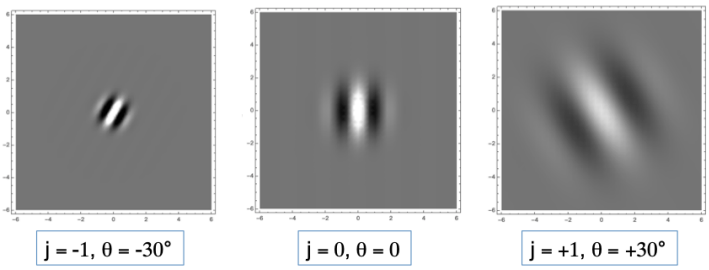

FIGURE 47 – Exemple d'ondelettes de la figure 46 dilatées et tournées.

les grandes structures. Ainsi, on définit des ondelettes tournées et dilatées selon [63]

$$\boxed{\psi_{\theta,j}(u) \equiv 2^{-2j}\psi(2^{-j}(r_{-\theta}.u))} \tag{139}$$

Notons que l'on a 1 ondelette par octave dans ce cas, et quelques exemples sont présentés sur la figure 47.

La question qui vient naturellement à l'esprit est comment choisir les angles? Tout d'abord, il faut se rendre compte que la transformée de Fourier de $\psi_{\theta,j}(u)$ est données par [64]

$$\widehat{\psi_\theta}(\omega) = \hat{\psi}(r_\theta.\omega) \tag{140}$$

Donc, l'idée des orientations et de bien couvrir l'anneau (Fig. 48) produit par les rotations successives donc on prend K rotations nécessaires. Lors d'un dilatation:

$$\boxed{\widehat{\psi_{\theta,j}}(\omega) = \hat{\psi}(2^j r_\theta.\omega)} \tag{141}$$

Voir la figure 48 pour une illustration de l'effet de rotation et dilatation dans l'espace de Fourier. Donc, on peut couvrir tout le plan de Fourier en choisissant bien la valeur de K.

63. NDJE: notez le scaling dimension n on aurait $1/s^n\psi(u/s)$.
64. NDJE en 2D, la TF fait intervenir $u^T\omega$, et r_θ est un matrice orthogonale $r_\theta^T = r_\theta$, $r_{-\theta}^{-1} = r_\theta$.

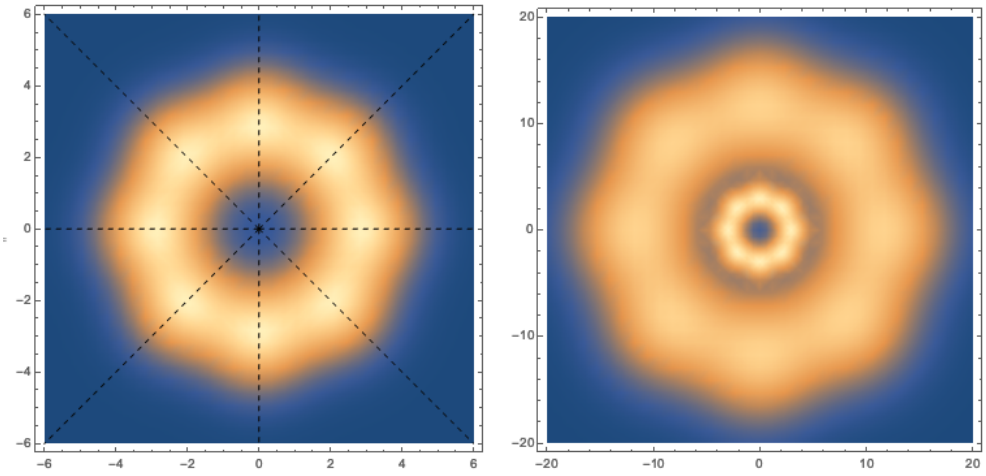

FIGURE 48 – Exemple de résultats dans le plan de Fourier, de rotations de d'ondelettes de la figure 46 par pas de 45° (gauche). A droite, on y ajoute la TF des ondelettes avec $j = -2$ (notez le changement d'échelles des axes).

On définit alors

$$\boxed{\psi_{j,k}(u) = 2^{-j}\psi(2^{-2j}r_{-\theta_k}.u) \qquad \theta_k = 2\pi\frac{k}{K}} \tag{142}$$

Ensuite, on définit la transformée en ondelettes selon le schéma développé en 1D

$$\boxed{Wx = (x * \psi_{j,k}, x * \phi_J)_{(j\geq J, 1\leq k\leq K)}} \tag{143}$$

avec ϕ_J le filtre basse-fréquence, cette fois en 2D, définie par la gaussienne g utilisée pour définir ψ, à savoir:

$$\phi_J(x) = 2^{-2J}g(2^{-J}x) \tag{144}$$

On a donc définit une transformée en ondelettes qui a explicitement utilisée le groupe des rotations qui se rajoute à la dilatation. La question qui vient alors est de savoir si W est un opérateur *complet et stable*? La réponse est oui à condition de **couvrir le plan de Fourier** sans avoir de trous. Ça se traduit par

$$1 - c \leq \sum_{k=1}^{K}\sum_{j=J}^{\infty} |\widehat{\psi_k}(2^j\omega)|^2 + |\hat{\phi}(2^J\omega)|^2 \leq 1 \tag{145}$$

et on en déduit comme pour en 1D que W est un opérateur contractant:

$$(1-c)||x||^2 \leq ||Wx||^2 = \sum_{k=1}^{K} \sum_{j=J}^{\infty} ||x * \psi_{j,k}||^2 + ||x * \phi_J||^2 \leq ||x||^2 \tag{146}$$

Finalement, **si on identifie x comme l'image d'entrée, l'action de W est de produire une collection de "canaux" indexés par j et k pour les $\psi_{j,k}$, additionnée du résultat du filtre basse fréquence, tout comme dans un réseau de neurones**.

Ensuite, on peut construire l'équivalent de MFC, c'est-à-dire un descripteur (représentation) $\Phi(x)$ selon

$$\boxed{\Phi(x) = \rho(x * \psi_{j,k}) * \phi_J = A_J \rho W x} \tag{147}$$

qui a les mêmes propriétés:

1. W produit de la **parcimonie** et il est **covariant/équivariant**,
2. ρ (ReLU) produit un opérateur **covariant/équivariant** qui n'est pas une simple moyenne,
3. le **pooling** (moyennage) ϕ_J dont le rôle est de créer un invariant qui élimine ce qui n'est pas nécessaire pour la classification.

Dans un réseau, on cascade ce genre d'opérations, qui plus est on apprend les W au lieu d'utiliser les ondelettes.

8.3 Quelques exemples

NDJE: S. Mallat présente des slides pour illustrer ce qu'il vient de développer sur le cas 2D. Je vous laisse les découvrir, ils sont présentés environ au bout de 30 minutes. Je vais donner les remarques sur lesquelles il insiste.

Les algorithmes rapides de la transformées en ondelettes (2D) travaillent par cascades de bancs de filtres (voir Cours 2018 Sec. 6.4). Chaque ondelettes $\psi_{j,k}$ est convoluée avec l'image originale, donc progressivement elle est translatée en 2D et le résultat n'est pas nul uniquement aux frontières/contours. En effet, par définition l'intégrale de ψ est nulle ce qui entraine que pour des zones constantes sur le support de ψ le résultat soit 0.

Les rotations permettent de détecter les frontières dans toutes les directions. Ainsi, les ondelettes manifestent la présence de variations locales (cf. comme les contours). A une échelle plus petite, en fait, on sous-échantillonne l'image d'origine, puis on réapplique les filtres, et ainsi de suite...

C'est le principe par exemple de la compression JPEG2000 des années 90, mais on voit réapparaitre ces cascades filtres/sous-échantillonnages dans les réseaux de neurones. Cependant, on a introduit le rectificateur ρ qui rend les coefficients positifs et donne de la puissance à basse fréquence qui interagit avec le filtre ϕ_J (comme en 1D). Jusque dans les années 2010 (cf. 2004-12), les descripteurs de type $\Phi(x)$ étaient construits "à la main (ex. DAISY [65]) c'est-à-dire sans recourir à l'apprentissage, et ils ont été la base de tous les algorithmes de reconnaissance du traitement d'images. Donc, on voit que cela soit en audio ou en imagerie, alors que cela a été fait indépendamment, le même type de descripteur a été élaboré et il est très efficace.

La question est Pourquoi? On a vu qu'il a été utile de rendre parcimonieuse la description du signal et utiliser les symétries du problème pour linéariser, d'utiliser une non-linéarité pour obtenir un invariant qui ne soit pas une simple moyenne qui perdrait toute la structuration du signal, et enfin le pooling pour réduire la dimensionalité. Mais le point qui reste à contrôler c'est **l'innocuité vis-à-vis des déformations**. Cependant, avant d'aborder ce point important, faisons un détour par la neurophysiologie.

8.4 Lien avec la neurophysiologie

Le lien avec l'audio a été abordé avec la rapide description de la cochlée et vous pouvez vous reporter au séminaire de Shihab Shamma. Dans l'imagerie, dans le cerveau il y a deux zones spécialisées dans le traitement (Fig. 38 et 49) situées à l'arrière du crâne. La zone "cortex visuel" dite V1, et "l'aire visuelle associative" dite V2. Fait extraordinaire, c'est que la zone V1 possèdent des **neurones sensibles à l'orientation** (Fig. 49) [66]. Ceci a été identifié entres autres dans les années 60 par **David Hubel** et **Torsten Wiesel** (Nobel 81)

65. NDJE: en butinant sur le Web, j'ai trouvé cette application qui décrit tous les descripteurs: https://tel.archives-ouvertes.fr/tel-01611384/document

66. NDJE: Source Michael C. Crair, Edward S. Ruthazer, Deda C. Gillespie, And Michael P. Stryker," Ocular Dominance Peaks at Pinwheel Centre Singularities of the Orientation Map in Cat Visual Cortex", Journal of Neurophysiology, vol. 77, 1997, p. 3381-3385.

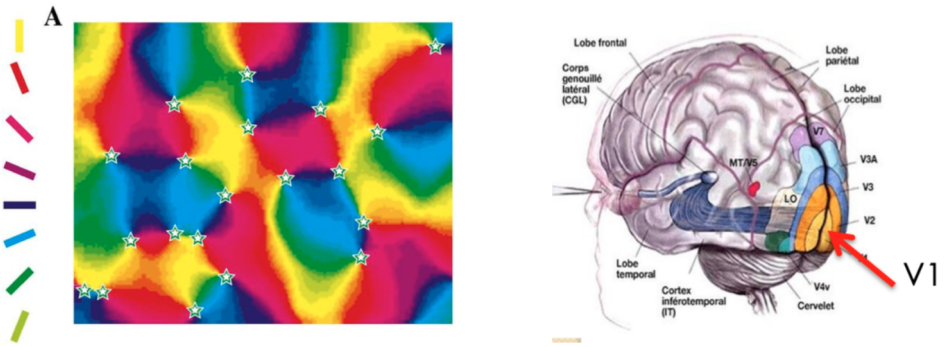

FIGURE 49 – A gauche: une image en fausse couleur montrant que les neurones de l'aire V1 du cerveau (image de droite) sont sensibles à l'orientation de motifs basiques: barre horizontale, verticale, oblique gauche et oblique droit.

qui partagèrent le prix avec **Roger W. Sperry** qui a quant à lui découvert la spécialisation fonctionnelles des deux hémisphères.

Hubel et Wiesel ont modélisé les neurones comme de simples filtres linéaires et ont mesuré la réponse impulsionnelle des neurones, à savoir le h dans l'opération $x * h$. Ils ont trouvé des **ondelettes** (de Gabor) avec des orientations spécifiques! Ensuite, ils ont remarqué qu'il y avait **des cellules plus complexes** dans l'aire V2: 1) elles sont très **non-linéaires**, 2) elles couvrent des **champs récepteurs beaucoup plus grands** et 3) surtout elles sont **invariantes par des transformations** notamment des translations, mais également des transformations beaucoup plus compliquées. Ensuite, dans V4 et IT (Fig. 49 droite) on a des réseaux de cellules **sensibles à des contours, des formes plus complexes et enfin à des objets, voire à des visages** (expérience anglaise de 2005) qui a donné lieu d'une manière humoristique à la « théorie du neurone grand-mère ». Selon cette théorie ce sont quelques neurones qui permettent de reconnaître le visage bien connu. C'est une théorie qui fait référence à la théorie computationnaliste (computation en anglais signifie calculable) en philosophie de l'esprit, originellement issue de Thomas Hobbes et qui répondait au formalisme chomskyen du langage [67].

La question est comment est-on passé des informations "orientationnelles" en V1 à des représentations très complexes et invariantes selon des transformations complexes en IT?

67. NDJE: j'ai reformulé selon mes recherches que j'avais effectuées pour un cours à des étudiants en Philosophie de la Nature.

Il y a des modélisations à l'aide de réseaux de neurones profonds, mais la question reste la même: comment passe-t'on des coefficients d'ondelettes (V1) à des descripteurs $\Phi(x)$ invariants (IT).

8.5 Stabilité par déformations

Derrière cette stabilité de $\Phi(x)$, on va s'apercevoir qu'il y a la propriété d'équivariance (covariance). En effet, on va montrer qu'une petite déformation va donner lieu à une petite translation, mais elle se fait à la fois dans l'espace et dans les échelles. Et encore une fois, la problématique est la même que vous fassiez de l'audio, de l'imagerie ou bien même de la chimie quantique: parcimonie, invariants par rapport des groupes de symétries du problème qui permettent d'éliminer des degrés de liberté, les déformations vont jouer un rôle dans la séparation des échelles avec des ondelettes, puis appliquer la non-linéarité et le pooling.

Théorème 10. *Il s'agit de démontrer la continuité Lipschitzienne par petite déformation, c'est-à-dire que*[68]

Soit une image x, et g une petite déformation qui lui est appliquée:

$$g.x(u) = x(u - \tau(u)), \text{ avec } ||\nabla\tau||_\infty < 1/2$$

Si on se restreint à des signaux $x \in L^2(\Omega)$ sur un support compact Ω[69] *qui peuvent être très irréguliers, alors:*

$$\begin{array}{l} \exists C > 0 \text{ tq. } \forall \tau \in C^2 \ ||\nabla\tau||_\infty < 1/2, \forall x \in L^2(\Omega), \\ \text{alors} \qquad \|\Phi(x) - \Phi(g.x)\| \leq C\|x\| \underbrace{\left(\underbrace{||\nabla\tau||_\infty + \|H\tau\|_\infty}_{|g|_G} + \underbrace{\|\tau\|_\infty 2^{-J}}_{\text{translation}} \right)}_{\text{taille de la déformation}} \end{array} \tag{148}$$

68. La condition porte sur la norme de la matrice jacobienne en dimension 2, ... en 1d il s'agit de simplement que $|\tau'(u)| < 1$. Ce faisant même si on peut aller jusqu'à 1, pour enlever toute instabilité autour de 1, on fixe la borne à 1/2. L'important est d'assurer l'inversibilité du difféomorphisme.

69. On peut étendre à tout $\mathbb{R}$ mais çà entraine des complications techniques

avec $\|H\tau\|_\infty$ *la norme infinie du Hessien (cf. les dérivées 2nd), et* $\|\tau\|_\infty = \sup_u|\tau(u)|$. *Notons que la norme de* $\Phi(x)$, *comme c'est un vecteur, on a*

$$\|\Phi(x)\|^2 = \sum_j \|\rho(x * \psi_j) * \phi_J(t)\|^2$$

Démonstration 10. Nous allons plutôt montré la structure de la démonstration. Ce que l'on veut c'est donc contraindre $\|\Phi(x) - \Phi(g.x)\|$, on a (on suppose que ϕ_J est toujours positive, ex. une gaussienne):

$$\begin{aligned}\|\Phi(x) - \Phi(g.x)\| &= \|\rho(W(x)) * \phi_J - \rho(W(g.x)) * \phi_J\| \\ &= \|(\rho(W(x)) - \rho(W(g.x)) * \phi_J\| && (* \text{ op. linéarire}) \\ &\leq \|(W(x) - W(g.x)) * \phi_J\| && (\rho \text{ contractant et } \phi_J \geq 0) \quad (149)\end{aligned}$$

On définit la transformée en odelettes selon la forme $Wx(u, \log s) = x * \psi_s(u)$ sachant que $s = a^j$. Donc, l'idée est de se demander si

$$g.Wx(u, \log s) = Wx(u - \tau(u), \log s) \overset{?}{\sim} W(g.x) \tag{150}$$

L'égalité signifierait que W et g commutent. Ce n'est pas tout à fait le cas mais considérons le lemme suivant:

Lemme 1.

$$\boxed{\begin{aligned}&\exists C > 0 \text{ tq. } \forall \tau \in C^2 \ ||\nabla\tau||_\infty < 1/2, \forall x \in L^2(\Omega), \\ &\qquad\qquad \|W(g.x) - g.W(x)\| \leq C\|x\| \, (||\nabla\tau||_\infty + \|H\tau\|_\infty)\end{aligned}} \tag{151}$$

C'est en fait le calcul d'un commutateur $[Wg - gW]$, si W est équivariant par l'action de g alors le résultat est nul. En fait la borne est gouvernée par la taille de la perturbation (cf. $|g|_G$).

Si on admet ce lemme, la démonstration du théorème 10 devient claire. Reprenons le

calcul Eq. 149, il vient

$$\begin{aligned}
&\|(W(x) - W(g.x)) * \phi_J\| \\
&\quad = \|(W(x) - g.W(x)) * \phi_J + (g.W(x) - W(g.x)) * \phi_J\| \\
&\quad \leq \underbrace{\|(W(x) - g.W(x)) * \phi_J\|}_{[Wg-gW]=0} + \underbrace{\|(g.W(x) - W(g.x)) * \phi_J\|}_{[Wg-gW]\neq 0} \\
&\quad \leq \|(W(x) - g.W(x)) * \phi_J\| + \|(g.W(x) - W(g.x))\| \qquad (\phi_J \text{ contractant}) \\
&\quad \leq \|(W(x) - g.W(x)) * \phi_J\| + C\|x\| \, (||\nabla\tau||_\infty + \|H\tau\|_\infty) \qquad (\text{lemme } 1)
\end{aligned} \tag{152}$$

La partie qui reste à traiter concerne ce qui reste après moyennage par ϕ_J de la différence entre une transformée en ondelettes ($W(x)$) et sa transformation par la déformation ($g.W(x)$). On va voir que ce qui reste c'est la translation globale. Soit le lemme suivant

Lemme 2. *Soit un signal z,*

$$\boxed{\exists C > 0 \text{ tq. } \|(z - g.z) * \phi_J\| \leq C' 2^{-J} \|\tau\|_\infty \|z\|} \tag{153}$$

En admettant ce lemme, il vient du fait que W est contractant

$$\|(W(x) - W(g.x)) * \phi_J\| \leq C' 2^{-J} \|\tau\|_\infty \|x\| + C\|x\| \, (||\nabla\tau||_\infty + \|H\tau\|_\infty) \tag{154}$$

ce qui donne le théorème si l'on prend le max entre C et C'.

Si l'on reprend les étapes de la démonstration, elle consiste à observer qu'**une déformation commute *presque* avec la transformée en ondelette**. Et l'erreur vient d'une part de la taille de la déformation, et d'autre part du résultat de la translation après moyennage qui est presque 0 à un terme d'erreur qui dépend de la taille de la translation par rapport à la taille du moyennage. Assurément quand la taille du moyennage par ϕ_J tend vers l'infini (cf. $J \to \infty$) alors ce terme disparait.

Donc il reste à démontrer les deux lemmes (1 et 2), dont le premier est difficile (mais intéressant). Les démonstrations complètes sont dans un article sur le site web [70]. S. Mallat

70. Voici le lien https://www.di.ens.fr/~mallat/College/TPAMI-Mallat-Bruna-Scat-CNN.pdf

nous propose de ne pas faire la démonstration complète de l'article car il la juge très longue, mais il nous propose de faire un développement de Taylor qui explique le phénomène de translation dans le domaine des échelles et pourquoi cela est très important.

Propriété: Si l'on prend le coefficient d'ondelette d'un signal déformé à une position u et une échelle $\log s$, alors

$$W(g.x)(v, \log s) \simeq Wx(v - \tau(v), \log s - \tau'(v)) \tag{155}$$

c'est-à-dire que **la déformation induit une translation à la fois dans le domaine spatial et dans le domaine des échelles logarithmiques**. Notons que la translation selon l'axe $\log s$ (cf. changement de l'indice car $\log s = i \times \log a$), va pouvoir être absorbée par la suite par convolution dans une étape ultérieure. Donc, voyons d'où cela vient

$$W(g.x)(v, \log s) = (g.x) * \psi_s(v) = \int x(u - \tau(u)) \frac{1}{s} \psi \left(\frac{v - u}{s} \right) du \tag{156}$$

Ensuite, on fait le changement de variable $u' = u - \tau(u)$ et tenir compte du fait que l'ondelette $\psi(x)$ est localisée autour de $x = 0$, donc on développe autour de $u = v$, cela donne (ici en 1D):

$$\tau(u) \simeq \tau(v) + (u - v)\tau'(v) \rightarrow u' = u - \tau(v) - (u - v)\tau'(v) \tag{157}$$

donc si $\tau(v)$ et $\tau'(v)$ petits, il vient

$$\begin{aligned} W(g.x)(v, \log s) &\simeq \int x(u') \psi \left(\frac{v - \tau(v) - u'}{s(1 - \tau'(v))} \right) \frac{du'}{s(1 - \tau'(v))} \\ &\simeq Wx(v - \tau(v), \log s - \tau'(v)) \end{aligned} \tag{158}$$

où l'on reconnait la convolution de x avec une ondelette translatée en $v - \tau(v)$ et rescalée en $\log s' \simeq \log s - \tau'(v)$. Une fois que l'on a ce résultat, on aimerait enlever la partie sur les échelles (cf. Eq. 150). Pour ce faire on va procéder à un développement de Taylor sur W:

$$W(g.x)(v, \log s) \simeq Wx(v - \tau(v), \log s) - \tau'(v) \left(\frac{\partial Wx(v - \tau(v), \log s)}{\partial \log s} \right) \tag{159}$$

Il nous faut donc connaître la sensibilité de la transformée en ondelettes le long de l'axe des échelles. Pour ce faire nous allons utiliser le lemme suivant (en 1D):

Lemme 3.

$$\boxed{\begin{aligned} \frac{\partial Wx(u,\log s)}{\partial \log s} &= -Wx(u,\log s) - \overline{W}x(u,\log s) \qquad (160)\\ \overline{W}x(u,\log s) &= x * \bar{\psi}_s \quad \text{avec} \quad \bar{\psi}(u) = u\psi'(u) \qquad (161)\end{aligned}}$$

qui se démontre facilement en repartant de la définition de la transformée en ondelette à une position u et échelle s, puis en procédant à la dérivée logarithmique qui fait apparaître la nouvelle ondelette $\bar{\psi}$. Le terme d'erreur devient alors

$$\begin{aligned} \|W(g.x)(v,\log s) - Wx(v-\tau(v),\log s)\| &\leq \|\tau'\|_\infty \|Wx + \overline{W}x\| \\ &\leq \|\tau'\|_\infty \left(\|x\| + \|\overline{W}x\|\right) \end{aligned} \qquad (162)$$

Or, l'ondelette $\bar{\psi}$ satisfait une condition de contractance à une constante prés et donc

$$\|W(g.x)(v,\log s) - Wx(v-\tau(v),\log s)\| \leq C\|\tau'\|_\infty \|x\| \qquad (163)$$

Ce qui démontre le lemme 1 sauf que dans le cas ici, on a tronqué la série de Taylor à l'ordre 1 pour donner l'idée. **Donc, qualitativement, quand on applique une déformation, le résultat est une translation à la fois en espace et en échelle, or l'erreur est de l'ordre de la translation selon les échelles qui est** τ'. Une remarque: la "vraie" démonstration ne procède pas du tout par un développement de Taylor qui a le problème entre autre du contrôle des ordres supérieurs. Le chemin est très différent mais autrement plus compliqué. Les détails sont dans l'article ci-dessus mentionné. Cependant, il n'en reste pas moins que, l'idée pour laquelle cela marche est bien celle qui vient d'être décrite.

Reste alors le lemme 2 dont nous allons donner l'argument qualitatif principal. On doit contrôler

$$(g.z) * \phi_J - z * \phi_J = \int z(u-\tau(u))\phi_J(v-u)du - \int z(u)\phi_J(v-u)du \qquad (164)$$

et comme $\|z\|^2 = \int |z(v)|^2 dv$, il vient après un changement de variable $u' = u - \tau(u)$ et

en laissant tomber les termes en $\tau'(u)$

$$\|(g.z) * \phi_J - z * \phi_J\|^2 = \int \left| \int z(u) \left(\phi_J(v-u-\tau(u)) - \phi_J(v-u)\right) du \right|^2 dv \qquad (165)$$

L'estimation de la différence du noyau ϕ_J en $v-u-\tau(u)$ et $v-u$ conduit à la contrainte que cette différence est majorée par la valeur sup. de la dérivée ϕ'_J, multipliée par la valeur sup. de la translation $\tau(u)$ le tout sur le support de ϕ_J. Ainsi

$$\begin{aligned} \|(g.z) * \phi_J - z * \phi_J\|^2 &\leq \int \int |z(u)|^2 | \left(\phi_J(v-u-\tau(u)) - \phi_J(v-u)\right) |^2 du dv \\ &\leq \int \int |z(u)|^2 ||\phi'_J||_\infty^2 ||\tau||_\infty^2 \mathbb{I}_{2^J}(v-u) du dv \end{aligned} \qquad (166)$$

La norme infinie de ϕ'_J fait sortir un 2^{-J}, ensuite l'intégrale qui reste sur u avec la l'indicatrice fit apparaître la norme de z. Ce qui finit de "démontrer" le lemme. ■

8.6 Bilan

Au delà de l'aspect technique de la démonstration, ce que l'on voit apparaître c'est la chose suivante: quand on a une forme de variabilité que l'on veut éliminer, alors on essaye de la transformer en une sorte de translation le long d'un nouveau paramètre, laquelle sera absorbée par moyennage. Dans le cas présenté jusque-là, le paramètre est l'échelle de dilatation s qui "absorbe" la déformation que l'on l'élimine par un pooling. Cependant, on perd de l'information, et c'est le véritable problème de tous ces descripteurs (MFC, DAISY...) qui ont été utilisés jusque dans les années 2010. Il est clair que les gens étaient bien conscient que ce moyennage était gênant, à savoir si on moyenne de trop, on perd la structuration originale qui permet de faire la classification; donc les moyennages ont été réalisés sur de petits domaines, ex. en audio 25 ms. Mais quand on veut capturer des structures sur des échelles beaucoup plus grandes on ne peut élargir le support de la moyenne, donc il faut utiliser d'autres techniques. Pour les images, le moyennage se faisait sur des patches de 8×8 pixels (voire maximum 16×16), mais les descripteurs locaux ne pouvaient pas capturer des structures bien au delà de ces tailles de patchs. Tout l'enjeu est donc de trouver des descripteurs qui sont sensibles aux grandes échelles. Pour faire cela, il faut cascader les transformations (transformation linéaire, ReLU, pooling), mais là

les choses deviennent plus compliquées à maitriser. Ce qui et surprenant néanmoins c'est que ces structures en cascade sont observées en neurophysiologie dans le cortex auditif et visuel, et dans les réseaux de neurones profonds.

9. Séance du 15 Juin

Cette séance est enregistrée après la période de confinement due à la COVID-19.

9.1 Quelques rappels

9.1.1 Les réseaux convolutionnels

S. Mallat tout d'abord nous redétaille rapidement l'architecture des réseaux convolutionels (Fig. 5). Les cascades de convolutions réduisent la dimensionalité de la représentation $\Phi(x)$ qui est par la suite traitée par un classificateur dense par exemple. La chose la plus étonnante est que ce genre d'architecture est capable de donner des résultats tout à fait remarquables dans des domaines aussi divers que la classification d'images, de sons, les différents aspects du traitement du langage, et également dans des problèmes de Physique/Chimie et Médecine. L'enjeu est vraiment de pouvoir **interpréter ses performances**.

Nous avons étudier 3 types de propriétés mathématiques pour faire apparaître les structures de ces réseaux de neurones: l'aspect **multi-échelles**, les **symétries** et la **parcimonie**. L'algorithmique des cascades de filtres et de sous-échantillonnage permet de visualiser que plus on s'enfonce dans les profondeurs de réseaux plus les neurones ont en charge de traiter des aspects à grandes échelles, par exemple de l'image initiale. Les symétries également sont un ingrédient essentiel du design de l'architecture et la première symétrie qui a été exploitée est **la covariance par translation** qui apparait notamment par l'usage de **convolutions**. La parcimonie (sparsity) apparaît si on peut dire comme un "accident" de l'apprentissage, et se traduit par le fait que sur une couche de neurones la réponse à un stimuli est essentiellement codée sur peu de neurones. Cependant cette parcimonie est fondamentale et nous aurons l'occasion d'y revenir dans le futur.

Quelles sont les questions:

— comment **la réduction de dimension** qui fait passer de $x \in \mathbb{R}^n$ à $\Phi(x) \in \mathbb{R}^d$ avec $d \ll n$ permet sans perdre d'information de répondre à la question $y = f(x)$?
— pourquoi les cascades **convolution/sous-échantillonage** sont-elles en quelque sorte efficaces pour répondre au problème?
— à quoi servent les **"non-linéarités"** ?
— la dernière couche $\Phi(x)$ en quelque sorte linéarise le problème, mais **que linéarise-t'on finalement?**
— dans les cascades de convolution, très tôt dans le réseau apparait un 3eme paramètre, "**le canal**": la question est d'interpréter le rôle de cette 3eme dimension, quel est l'objet mathématique qu'il y a derrière? On verra quelle permet d'exprimer des **notions de symétries** et sert également pour construire des **représentations parcimonieuses**.

9.1.2 Les symétries du problème

L'enjeu comme on le sait, est la représentation/**approximation d'une fonction f en très grande dimension**, ce qui pose problème car les points (exemples pour l'entrainement par ex.) sont très loin les uns des autres ce qui impose alors de trouver des sources de **régularité très fortes** sur la fonction f sous-jacente. Nous avons vu dans le cours de cette année (par ex. Sec. 5) comment étudier le groupe des symétries du système permet de trouver ces formes de régularité en très grande dimension. On peut voir un symétrie comme un opérateur qui transforme les éléments d'une classe en un autre élément de cette même classe. Prenons l'ensemble Ω_t des iso-valeurs de f définit par

$$\Omega_t = \{x, f(x) = t\} \tag{167}$$

Alors g est une symétrie de f à savoir

$$f(g.x) = f(x) \tag{168}$$

si g préserve les Ω_t. En effet, soit $x \in \Omega_t$ et soit g une symétrie de f alors: $f(g.x) = f(x) = t$ donc $g.x$ est un élément de Ω_t. Dans le cas d'une régression pour laquelle t prend des valeurs continues, g préserve *les lignes de niveaux* de f. L'ensemble des symétries g de f

a une structure de groupe [71]: c'est **le groupe des symétries de** f.

L'usage des symétries de f, se fait en fait à travers **les symétries de la représentation** Φ qui est apprise dans le cas d'un réseau de neurones. Il y a une autre façon qui vise à imposer directement la/les symétrie(s) en ayant un *a priori* sur les symétries du problème que l'on a vu à travers les descripteurs MFC à base d'ondelettes (voir Secs. 7 et 8), et la dernière phase (la linéarisation) est apprise pour s'adapter au problème posé. Notons que dans le cas d'un réseau de neurones, l'information *a priori* est contenue dans l'architecture du réseau [72]. Donc, il y a **deux étapes distinctes**: une linéarisation par l'action de groupe de x vers $\Phi(x)$, et une linéarisation de $\Phi(x)$ vers y la réponse du réseau.

La première symétrie couramment rencontrée est **la translation**. Cependant, on peut aller un cran plus loin en analysant des exemples tels que celui de la reconnaissance de chiffres. Il est clair qu'une petite déformation locale ne change pas la classe du chiffre. Et donc si x est un 3, alors x' définit par [73] une translation dépendante de la position selon

$$x'(u) = x(u - \tau(u)) \tag{169}$$

avec $||\tau|| \ll 1$, il s'avère que x' est aussi reconnu comme un 3 comme sur la figure 15. Donc, ces petites déformations qui constituent des **difféomorphismes** sont des symétries de ce problème. Or, ceci est très important comme on l'a vu car, **de la dimension du groupe sous-jacent dépend le facteur de réduction de la dimensionalité** du problème. Dans le cas des difféomorphismes, on tient là un **groupe de dimension colossale**, donc bien adapté à la réduction de la dimensionalité des images à plusieurs millions de pixels, et/ou des sons à plusieurs milliards d'échantillons. Cette réduction permet d'**appréhender des problèmes qui nécessitent de regarder soit l'image soit une phrase musicale/vocale dans leur entièreté**. Dans le cas de l'analyse des voies, les translations temporelles et les transpositions fréquentielles font passer d'un locuteur féminin à un locuteur masculin dont les timbres et rythmes sont différents.

Il y a d'autres types de symétries: par exemple **la rotation** associée ou non à des **déformations locales**. Par exemple, la reconnaissance de textures d'écorces d'arbres doit

71. Simplement, on étudie la composition $g_1.g_2$, et l'on note que $f(g_1.g_2.x) = f(g_2.x) = f(x)$.
72. Rappelons nous également que nous avons une compréhension du Monde également emprunte d'un *a priori* dû à l'architecture de notre cerveau
73. u est la variable de position par ex. du pixel en 2D

intégrer ce type de symétries du groupe $SO(2)$. Également, il est fréquent de rencontrer des **changements d'échelles** (avec/sans déformations) simplement par l'effet de zoom sur des structures par exemple. Dans ce cas le groupe est $\mathbb{R}$. Et donc comme on peut le remarquer, les symétries du problème que l'on veut préserver dans la représentation $\tilde{f}$ de la fonction f solution, sont très variées et dépendantes du problème posé (reconnaissance d'images ou de voies). On se rend compte aussi qu'il y a beaucoup de symétries que l'on ne connait pas, et que la phase d'apprentissage d'un réseau permet d'appréhender.

9.1.3 Création/utilisation des invariants

Rappelons les étapes: à partir de x en entrée, on fixe/apprend une représentation $\Phi(x)$ et pour obtenir une approximation $\tilde{f}(x)$ de la fonction $f(x)$ sous-jacente, on linéarise selon (voir Sec. 5.4)

$$\tilde{f}(x) = \langle w, \Phi(x) \rangle = \sum_k w_k \phi_k(x) \tag{170}$$

(nb. rappelons que w est appris même si dans la première phase on se donne Φ). Si on veut garantir que les symétrie de f sont préservées c'est-à-dire que

$$\forall g \in G,\ \tilde{f}(g.x) = \tilde{f}(x) \tag{171}$$

alors on veut que

$$\forall x,\ \langle w, \Phi(x) - \Phi(g.x) \rangle = 0 \tag{172}$$

c'est-à-dire que w et $\Phi(x) - \Phi(g.x)$ soient orthogonaux. Nous avons alors deux stratégies selon le cas de figure:

- **soit G est connu**, alors on va essayer de trouver Φ telle que $\Phi(x) = \Phi(g.x)$, c'est-à-dire que Φ est équivariante par rapport aux éléments de G. Cependant, à part des problèmes à petites dimensions ou très spécifiques, bien souvent on ne connait que partiellement le groupe de symétries.
- soit $G \subset G'$ avec G connu mais G' **inconnu**. Alors, on essaye de garantir qu'il existe un w qui définit la normale à un hyperplan dans lequel $\Phi(x) - \Phi(g.x)$ évolue avec $g \in G'$ (voir Fig. 17). Par ce mécanisme, **on "tue" la variabilité du problème selon l'action de** G (nb. il reste la variabilité sur les autres symétries potentiellement cachées que l'on pourrait tenter de mettre en évidence dans une post-analyse). La direction w est apprise dans la dernière couche linéaire (partie dense) du réseau.

Donc, pour apprendre l'action de groupe $g \in G'$, on peut linéariser $\Phi(x) - \Phi(g.x)$ en utilisant des petites transformations, et si on impose que Φ est Lipschitz (forme de régularité) alors (voir les développements qui conduisent à Eq. 46)

$$\|\Phi(x) - \Phi(g.x)\| \leq C|g|_{G'}\|\Phi(x)\| \tag{173}$$

c'est-à-dire que **la différence entre $\Phi(x)$ et $\Phi(g.x)$ doit être de l'ordre de la transformation** g. Si Φ est "dérivable" (au sens faible) par rapport à l'action de g alors le plan tangent existe.

On a vu comment ce schéma pouvait être mis en œuvre dans les cas suivants:
— celui les translations $g.x(u) = x(u - \tau)$ alors on impose que $\Phi(g.x) = \Phi(x)$
— celui des difféomorphismes où

$$g.x(u) = x(u - \tau(u))$$

expression qui peut être linéarisé, et on trouve alors que la transformation se décompose en une *translation globale* et une *déformation locale* telles que (voir Eq. 43)

$$|g|_G = \|\tau\|_\infty + \|\nabla\tau\|_\infty$$

9.2 Mise en application dans un réseau de neurones

Nous allons montrer que nous pouvons mettre en application le programme de prise en compte des symétries *a priori* du problème dans un réseau dont les filtres sont complètement connus et à base d'ondelettes (**réseau de Scattering**). C'est-à-dire que **dés lors que l'on connait le groupe de symétrie du problème, on a pas besoin d'apprendre les filtres.** En particulier, c'est vrai pour les symétries telles que: translation, rotation, transposition fréquentielle, ou même des déformations.

L'architecture à laquelle on aboutit est schématiquement celle de la figure 50. Elle ressemble à s'y méprendre à celle d'un réseau convolutionnel tel qu'on le conçoit couramment de nos jours (cascade de filtres/sous-échantillonnage et non-linéarité). Une différence notable est que l'on a plutôt ici des structures en arbres d'une couche à l'autre de filtres (voir les L_j colorés) et dans un premier temps **on ne fait pas communiquer les différents**

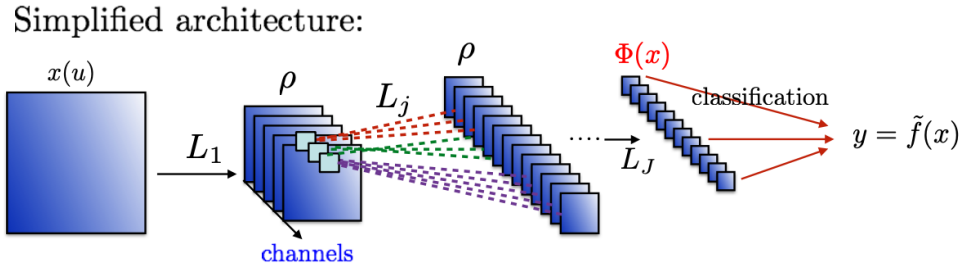

FIGURE 50 – Représentation graphique d'une architecture purement basée sur des filtres en Ondelettes.

canaux (cf. le long du 3eme axe) ce qui va simplifier le problème. Cependant, ce que l'on va apprendre/constater c'est que l'on va aboutir à **des réseaux dont les performances sont limités et moindre que celle des CNN 'classiques'**. On va ainsi apprendre ce que **ces canaux** apportent dans la résolution du problème et pourquoi ils rendent si efficaces les CNN 'classiques'.

9.3 Étape 1: séparation d'échelle

On va utiliser la Transformation en Ondelettes que l'on a abordé en 2018 et détaillé durant cette année plus particulièrement. Mettons que le problème soit relié à l'analyse de signaux qui dépendent de la variable temps t, on va utiliser une ondelette $\psi(t)$ que l'on translater et dilater par une échelle λ:

$$\psi_\lambda(t) = \lambda^{-1}\psi(\lambda^{-1}t) \tag{174}$$

ans le cas de l'audio on prend $\lambda = 2^{j/Q}$. La base d'ondelettes est formée par les dilatations et les translations. La Transformée en Ondelettes consiste à calculer la corrélation du signal x avec l'ondelette dilatée, ce qui se traduit par la convolution/filtrage

$$(x * \psi_\lambda)(t) = \int x(u)\psi_\lambda(t-u)du \tag{175}$$

Dans le domaine de Fourier, cette convolution se traduit par le produit des TF à savoir

$$\widehat{(x * \psi_\lambda)}(\omega) = \widehat{x}(\omega)\ \widehat{\psi_\lambda}(\omega) \tag{176}$$

On a vu dans le cours des exemples d'ondelettes et dans le domaine de Fourier, $|\widehat{\psi_\lambda}|^2$, se sont des **filtres passe-bandes** (voir Fig. 40).

Si on comprime l'ondelette dans le domaine temporelle, on la dilate et translate vers les hautes fréquences dans le domaine de Fourier. En échelle logarithmique la largeur du filtre de Fourier est constante (voir Fig. 39) ce qui donne lieu à l'appellation de "*Q-constant band-pass filters*". On a vu que si on coupe à basse fréquence la décomposition à une échelle 2^J, on doit associé un **filtre passe-bas** ϕ_{2^J} **qui va donner une approximation moyenne** de la fonction, tandis que les **Q-filtres vont donner les détails** de la fonction, et donc la décomposition complète en Ondelettes se fait selon ($\lambda = a^j$ avec $a = 2^{1/Q}$)

$$Wx = \begin{pmatrix} x * \phi_{2^J} \\ \\ x * \psi_{a^j} \end{pmatrix}_{a^j \leq 2^J} \tag{177}$$

Si les ondelettes couvrent bien la bande de fréquences alors la transformation préserve la norme du signal, cf. $\|Wx\|^2 = \|x\|^2$.

Un exemple de décomposition est donné sur la figure 51. Nous avons vu également dans la section 7.3 et les séminaires de Geoffroy Peters du 12 février 2020 et de Shihab Shamma du 12 Mars 20, que ces types de décompositions en ondelettes sont implémentées physiologiquement notamment dans *l'appareil auditif de la cochlée* et dans le *cortex auditif*, où non seulement de fortes non-linéarités apparaissent mais aussi que les réponses des neurones sont insensibles à la translation, et donc permettent traiter des informations à grande échelle de temps. Bien entendu, on aimerait bien comprendre les traitements qui se font dans le cortex auditif (de même pour le cortex visuel).

Dans le cas *des images*, les principes sont essentiellement les mêmes. Par exemple, l'ondelette peut être représentée par une fonction complexe comme une gaussienne modulée par un cosinus pour la partie réelle et un sinus pour la partie imaginaire. Cependant, à la translation et dilatation, on y ajoute les rotations. Ainsi, une ondelette transformée s'écrit selon (voir Sec. 8.2, Fig. 47)

$$\psi_{\theta,j}(u) = 2^{-2j}\psi(2^{-j}(r_{-\theta}.u)) \tag{178}$$

Un exemple est donné sur la figure 52 qui montre que de telles ondelettes sont sensibles

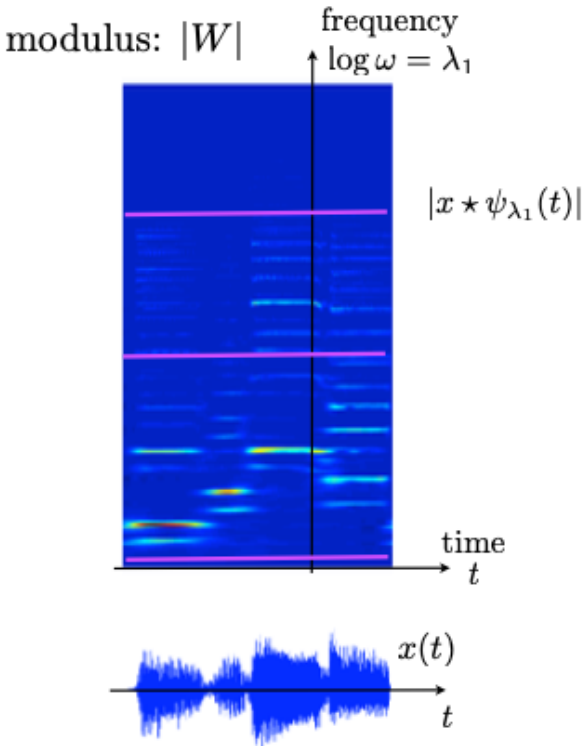

FIGURE 51 – Exemple de décomposition en Ondelettes d'un signal $x(t)$: on représente en couleur la valeur de $\|Wx\|$ avec le bleu représentant des valeurs nulles. On se rend compte de la parcimonie de la décomposition en fréquence et de son évolution temporelle.

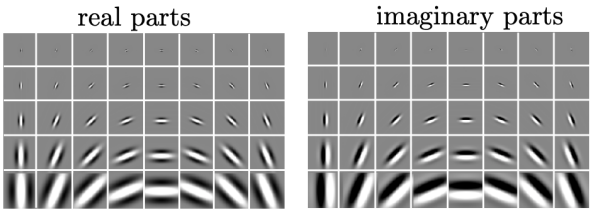

FIGURE 52 – Représentation d'ondelettes complexes dilatées et tournées.

à l'orientation des détails de l'image. La transformée en ondelettes tout comme en audio, est un filtrage de l'image par convolution

$$(x * \psi_{\theta,j})(u) = \int x(v)\psi_{\theta,j}(u-v)dv \tag{179}$$

que l'on peut visualiser également dans le domaine "fréquentiel" comme sur la figure 48. On peut ensuite, ajouter une fonction ϕ_{2^J} qui couvre les "basses fréquences" et on définit Wx sur le même principe que l'équation 177 selon (notez qu'ici pour l'image $a = 2$)

$$Wx = \begin{pmatrix} x * \phi_{2^J} \\ \\ x * \psi_{2^j,\theta} \end{pmatrix}_{j \leq J,\theta} \tag{180}$$

avec les mêmes propriétés concernant la conservation de la norme du signal. Ce qui est intéressant (voir le séminaire de Simon Thorpe) c'est que ces ondelettes ont été mises en évidence dans le cortex visuel. De même, quand on rentre plus profondément dans les structures du cerveau, on constate tout comme pour l'audio des fortes non-linéarités et des formes d'invariance par translation d'abord, et ensuite que l'on rencontre de petits réseaux sensibles à des effets très globaux qui rendent par exemple la reconnaissance d'un visage invariant par tout un tas de changements de positions/texture/vieillissement/etc. Comprendre comment le cerveau traite ces informations aux différents stades est un sujet de recherche actuel en lien avec les progrès de la neurophysiologie.

D'un point de vue mathématique pourquoi les ondelettes sont-elles efficaces et semble-t'il à l'œuvre dans le cerveau? Revenons à des arguments traités cette année:

— tout d'abord, il y a une forme de **stabilité par rapport à des déformations**. Si on

opère la transformation suivante

$$\psi_\lambda(u) \to \psi_\lambda(u - \tau(u)) = \psi_{\lambda,\tau}(u)$$

alors l'écart entre l'ondelette transformée et la copie originale est de l'ordre de la déformation

$$\|\psi_\lambda - \psi_{\lambda,\tau}\| \leq C\|\nabla\tau\|_\infty$$

— on a naturellement une **séparabilité des échelles** avec la transformation en ondelettes qui aide à la réduction de dimensionalité (voir Sec. 4.3);
— enfin la transformation en ondelettes fournie une représentation très **parcimonieuse** (sparsité) qui permet de mettre en évidence *les features* du signal.

9.4 Étape 2: invariants par translation

Les descripteurs doivent être capables de capturer ce type d'**invariance par translation**. Les coefficients en ondelettes aident à la construction de ces descripteurs comme on l'a vu pour les MFC (audio) et leurs équivalents pour le traitement des images.

En fait, on n'a pas vraiment de choix: **il faut moyenner le signal d'une certaine façon**. Si on veut obtenir une moyenne à une échelle 2^J, on procède à la convolution du signal x par la fonction "basse-fréquence" ϕ_{2^J}. Le résultat est une approximation de f qui par translation donne des coefficients en ondelettes qui ne varient pas beaucoup. A l'extrême, si on impose vraiment l'invariance par translation, alors il faut complètement moyenner le signal, ce qui se fait quand $2^J \to \infty$; mais dans ce cas toute la structuration du signal est perdue. Ce que l'on a vu dans le cours cependant c'est que **les non-linéarités** sont nécessaires pour obtenir des invariants qui capturent des informations "perdues" par le moyennage.

Les informations "perdues" sont les variations hautes-fréquences du signal que l'on peut capturer par la transformation en ondelettes en utilisant les ψ_λ pour différentes valeurs de λ (nb. ici λ est une échelle générique adaptée pour le cas de l'audio ou de l'image). Cependant, comme ψ est une fonction oscillante de moyenne nulle, **si on se contente de moyenner les coefficients en ondelettes du signal, on obtient une valeur nulle**. Donc, on

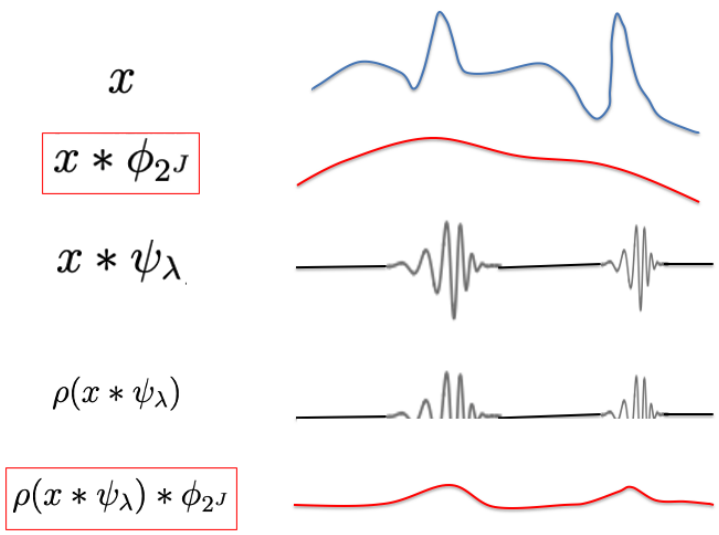

FIGURE 53 – Élaboration d'invariants par translation (en rouge) à partir des coefficients en ondelettes du signal en utilisant le rectificateur ρ pour la partie haute-fréquence, et le moyennage à l'aide de ϕ_{2^J}.

ne peut utiliser un filtre linéaire. Par contre, on peut utiliser un rectificateur [74] ρ qui met les coefficients à zéro s'ils sont négatifs. **Ainsi, moyenner $\rho(x * \psi_\lambda)$ ne donne plus une valeur nulle, et permet de préserver l'information de localisation des variations rapides du signal**. Pour construire alors un invariant à partir des coefficients rectifiés, on procède par moyennage avec ϕ_{2^J} comme pour le signal lui-même: on obtient alors $\rho(x * \psi_\lambda) * \phi_{2^J}$. Les différentes étapes d'élaboration d'invariants par translation qui ne perdent pas la structuration du signal sont schématisées sur la figure 53.

9.5 Opérateurs de Scattering

Ce qui reste néanmoins vrai dans les étapes décrites ci-dessus c'est que l'on fait des moyennes sur une échelle 2^J qui potentiellement peut-être grande, donc on perd qu'en même quelque chose. Nous avons vu que de tels descripteurs (MFCC, SIFT) très utilisés (avant les réseaux convolutionnels) ont l'inconvénient majeur de n'être sensibles qu'à des échelles somme toute assez petites. Pourquoi? la raison en est que si le support de ϕ_{2^J}

74. Notons que l'on peut utiliser le module des coefficients.

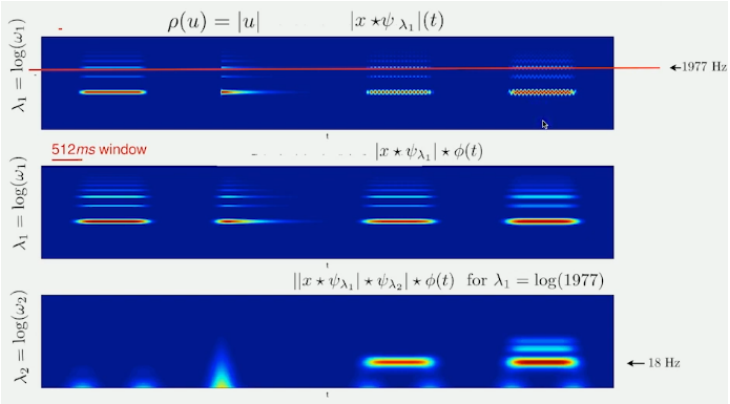

FIGURE 54 – Exemple du résultat de la cascade d'opérateurs de scattering. Le scalogramme du haut est celui du signal composé de trois sons de gauche à droite: une note simple, une "attaque", un trémolo et un vibrato. Le premier étage (scalogramme du milieu) qui calcule $\rho(x * \psi_{\lambda_1}) * \phi_{2^J}$ va garder les structures en harmoniques mais perd dans la structuration: on ne voit pas vraiment l'attaque et les ondulations du trémolo et vibrato sont perdues. Le second étage (scalogramme du bas) qui effectue $\rho(\rho(x * \psi_{\lambda_1}) * \psi_{\lambda_2}) * \phi_{2^J}$ par exemple pour λ_1 correspondant à 1977 Hz donne des coefficients peu intenses pour la première note, montrent des coefficients assez grands pour marquer l'attaque de la seconde note, et donnent des coefficients grands exactement aux fréquences du trémolo et du vibrato.

est trop grand on perd de l'information, mais s'il est trop petit on perd l'invariance par translation. Or, on aimerait pouvoir obtenir des invariants sur des domaines bien plus larges. En fait, on peut retrouver les hautes fréquences des coefficients rectifiés, cf. $\rho(x * \psi_{\lambda_1}) * \psi_{\lambda_2}$ que l'on va de nouveau rectifier puis moyenner... **On sent bien la nécessiter d'enchaîner/de cascader des séries de couples "filtrage/rectification".** Ces cascades sont appelés des **opérateurs de scattering** qui sont une version simplifiée des réseaux de neurones. Un exemple complet est donné sur la figure 54 avec 4 types de sons ayant des structurations différentes. **A chaque étage de scattering on retrouve les informations qui ont été gommées à l'étage antérieur.**

En quoi, cette architecture est reliée à celle des réseaux de neurones? Dans un réseau de neurones, on applique une succession de convolutions et de non-linéarités. Ce que l'on va illustrer,s c'est que ces successions de convolutions précisément mettent en jeu une décomposition en ondelettes. Pour ce faire, prenons un signal $x(u)$ en 1D, et donnons nous comme objectif de calculer la moyenne de ce signal. Bien entendu, on pourrait faire

le calcul directement. Cependant, on peut remplir cet objectif efficacement en agrégeant les valeurs par paires. On obtient alors successivement

$$\begin{aligned} x_0(u) &= x(u) \\ x_1(u) &= H[x_0](u) = \frac{x_0(2u) + x_0(2u+1)}{2} \\ x_j(u) &= H[x_{j-1}](u) = H^{(j)}[x_0](u) \\ &\dots \end{aligned}$$

Donc, si $x(u)$ à 2^N valeurs initiales, on obtient successivement 2^{N-1}, 2^{N-2}, etc valeurs intermédiaires, et au final 1 seule valeur en N étapes, laquelle est la moyenne du signal. Donc obtenir une information invariante par translation est très simple.

Mais à chaque étape, on perd de l'information. Ici, l'information perdue correspond aux différences par paire. On a donc un premier filtre H pour les moyennes 2-à-2, et un second filtre G pour les différences 2-à-2:

$$\{x(u)\}_{u\leq d} \rightarrow \begin{cases} \left\{\frac{x(2u)+x(2u+1)}{\sqrt{2}}\right\}_{u\leq d/2} & (H) \\ \left\{\frac{x(2u)-x(2u+1)}{\sqrt{2}}\right\}_{u\leq d/2} & (G) \end{cases} \tag{181}$$

Les deux opérations Hx et Gx sont en fait deux convolutions avec pour la première un filtre passe-bas h (valeurs $(1,1)$) et la seconde un filtre passe-bande g (valeurs $(1,-1)$):

$$Hx(u) = x * h(2u) \qquad Gx(u) = x * g(2u) \tag{182}$$

Si on itère la décomposition sur les approximations successives (H,$H^{(2)}$,$H^{(3)}$, ...) on obtient une arbre comme sur la figure 55 qui fait apparaître l'ondelette ψ de Haar et la fonction de scaling associée ϕ. Bien entendu ce schéma de cascade de filtres passe-bas et passe-bande est généralisable au delà de l'usage de l'ondelette de Haar. Une version en 2D est illustrée sur la figure 56 où l'on remarque encore une fois le caractère parcimonieux de la décomposition: presque tous les coefficients sont nuls sauf aux frontières entre surfaces de même intensité de gris dans l'image initiale.

Maintenant comme nous l'avons vu, une fois la décomposition en ondelettes, afin de trouver des invariants on doit dans un premier temps appliquer un opérateur non-linéaire

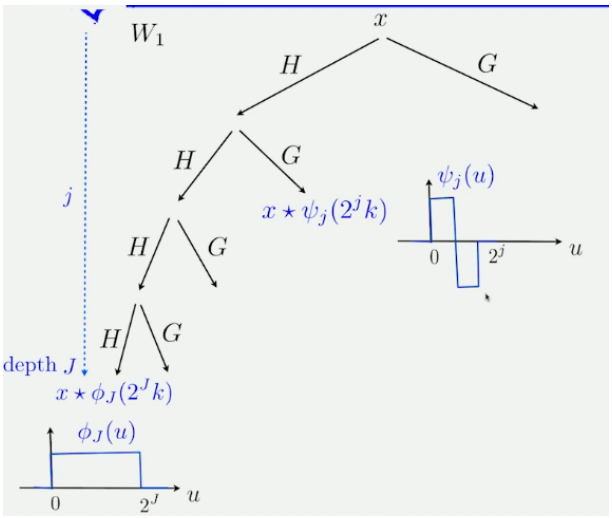

FIGURE 55 – Cascade successives de l'application des filtres passe-bas (H) et passe-bande G qui fait apparaitre l'ondelette ψ de Haar et la fonction de scaling associée ϕ.

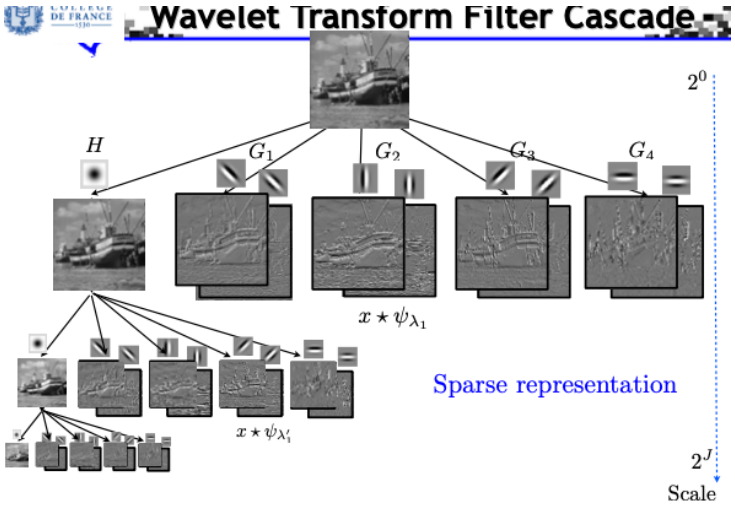

FIGURE 56 – Même type de cascade que pour la figure 55 avec deux types de filtres passe-bas et passe-bande dans le cas d'un traitement d'image. Les zones sombres indiques des zones pour lesquelles les coefficients sont nuls.

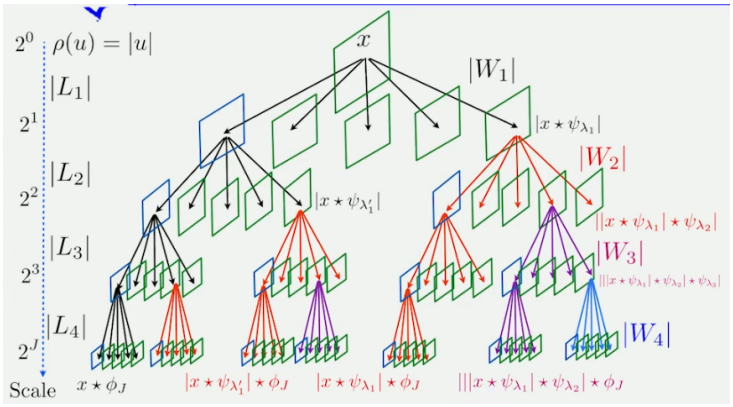

FIGURE 57 – Décomposition en cascade d'une image: à chaque étape on obtient une décomposition en ondelettes dont on extrait un invariant par application d'une non-linéarité (ici le module) et par moyennage à l'échelle 2^J. On obtient alors successivement un invariant du 1er ordre $x * \phi_J$, des invariants du 2nd ordre $|x * \psi_{\lambda_1}| * \phi_J$, puis du 3eme ordre $||x * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \phi_J$, etc.

sur les coefficients de détails à chaque niveau d'échelle 2^j (un ReLU ou le module par exemple), et dans un deuxième temps procéder à une moyenne. Parallèlement, chaque image de détails peuvent également être analysées par une cascade d'ondelettes du second ordre, et ainsi de suite. On voit alors apparaître le schéma en arbre de la figure 57. L'interprétation de cet arbre de décomposition peut être donc celle d'**un réseau de neurones où tous les filtres sont des ondelettes** (scattering network):

$$S_J(x) = \{|||x * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \dots \psi_{\lambda_m}| * \phi_J\}_{\lambda_k} \tag{183}$$

dont horizontalement on a l'application de filtres W_k, et verticalement **on extrait des invariants de plus en plus globaux** (cf. plus on s'enfonce dans le réseau plus l'échelle d'introspection est grande).

Nous avons vu (cf. Th. 8) que cette décomposition en cascade de filtres, de non-linéarités et de moyennages est **contractante**. Pour mémoire,

— la décomposition en ondelettes est un opérateur linéaire qui conserve la norme, cf. $\|Wx\| = \|x\|$

— la non-linéarité est un opérateur contractant car $\forall a, b$, $|\rho(a) - \rho(b)| \leq |a - b|$, donc

$$\|\rho W(x) - \rho W(x')\| \leq \|x - x'\|$$

et donc comme $S_J(x)$ est constitué d'enchainement d'opérateurs contractants alors S_J**, la sortie du réseau de scattering, est un opérateur contractant** donc stable en norme L2:

$$\boxed{\|S_J(x) - S_J(x')\| \leq \|x - x'\|} \tag{184}$$

Autre résultat que nous avons étudié en détails (Sec. 8.5) c'est **la stabilité de** S_J **vis-à-vis de (petites) déformations**, c'est une régularité Lipschitzienne, et on a ce type d'inégalité

$$\boxed{\text{Si } D_\tau(x)(u) = x(u - \tau(u)), \text{ alors } \lim_{J \to +\infty} \|S_J.D_\tau(x) - S_J(x)\| \leq C\|\nabla\tau\|_\infty \|x\|} \tag{185}$$

(nb. ici pour simplifier on a omis le Hessien de la transformation). En fait cela est dû au fait qu'une déformation correspond à peu de chose prés à une translation le long des échelles.

Pourquoi est-ce important d'obtenir des invariants stables par déformations? Encore une fois, le raisonnement est le suivant: si le problème est invariant par telle ou telle transformation, incorporer cette information *a priori* dans la structure du réseau permet de linéariser le problème, ou en d'autres termes de s'affranchir de cette variabilité, rendant ainsi le problème d'approximation plus simple car de plus basse dimension. Et cela est d'autant plus efficace que le groupe d'invariance est de grande dimension, d'où l'importance des difféomorphismes.

9.6 Quelques applications des réseaux de Scattering

9.6.1 La classification des digits

Le premier cas d'étude est celui de **la classification des digits MNIST** réalisée par Joan Bruna Estrach [75] en 2012. Le réseau était constitué en deux blocs: le premier était

75. Voir sa thèse https://www.di.ens.fr/data/publications/papers/phd_joan.pdf et les articles afférents.

constitué du réseau de Scattering qui linéarisait les translations et les petites déformations (difféomorphismes) avec des filtres en ondelettes, et un second bloc spécifique à la classification des digits. Le premier bloc ne nécessite aucun apprentissage, le second bloc apprend le vecteur w de la classification par régression logistique (par ex.). Les résultats de l'époque donnait une erreur de classification de 0.4% identique à celle obtenue par le réseau convolutionnel de Y. LeCun et al [76]. Ceci dit dans le cas de MNIST, on dispose d'une base de données d'entrainement de 50,000 digits (et 10,000 pour le test), donc on en a suffisamment pour apprendre tous les filtres ainsi que les paramètres du classificateur final. Donc, rien d'étonnant finalement à ce que les résultats soient identiques surtout que l'essentiel est de pouvoir capturer la variabilité due aux translations et déformations, tâche pour laquelle les ondelettes sont intrinsèquement bien adaptées.

9.6.2 La classification des textures

Un autre problème est celui de **la classification de textures de la base CUReT** (Columbia-Utrecht Reflectance and Texture Database [77]) analysée par Joan Bruna Estrach également. **Il s'agit d'un petit lot de textures** de 200×200 pixels distribuées sur 61 classes. Là le lot d'entrainement est extrêmement petit par rapport à la tâche de classification, puisque que l'on ne dispose que de 46 images par classe. Indubitablement on ne peut mettre en œuvre un réseau tel que `AlexNet` ou du même type ayant de l'ordre de 60M de paramètres [78]. **L'intérêt du réseau de scattering est que la première partie du réseau n'a pas besoin d'être apprise** et l'apprentissage est une "simple" régression logistique à 61 paramètres. Le taux d'erreur obtenu est de 0.2% bien inférieur au résultat obtenu par analyse de Fourier qui donnait 1% d'erreur. Ceci est du au fait que deux textures différentes peuvent avoir le même spectre de puissance, donc pour les différentier **il faut pouvoir analyser les non-gaussianités** ce que fait le réseau de scattering [79].

76. Notons que les méthodes de PCA ou SVM n'obtiennent pas ce niveau d'erreur.
77. https://www.cs.columbia.edu/CAVE/software/curet/, et voir la thèse de Joan Bruna Estrach.
78. Notons cependant que la plupart des paramètres sont ceux de la partie dense du classificateur et non de la partie convolutionnelle.
79. S. Mallat indique que cela fera l'objet d'un cours: l'analyse des non-gaussianités avec ce genre de réseau.

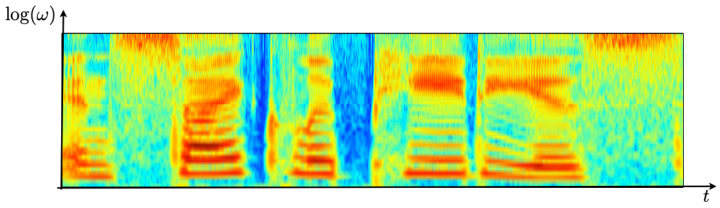

FIGURE 58 – Scalogramme d'une trame sonore: l'axe horizontal est le temps, et l'axe verticale l'échelle de la décomposition en ondelettes, c'est aussi l'axe des "canaux". L'échelle des couleurs indique l'intensité des coefficients d'ondelettes.

9.6.3 Le rôle des connexions entre canaux

Si on analyse le schéma type d'un réseau de scattering de la figure 57 (ou 50), on se rend compte de **l'absence** de l'opération de la figure 4 qui relie les images le long de **l'axe des canaux**. **Les résultats des filtres W_k ne sont pas connectés**. C'est la grosse différence par rapport à un réseau de neurones 'classique'. Donc, les questions qui viennent à l'esprit sont: **quel est le rôle de ces connexions?** et **pourquoi sont-elles importantes?** En fait nous allons voir quelles participent à la linéarisation **d'autres types de symétries** au delà des translations/déformations et quelles permettent d'**apprendre des patterns discriminants**.

Revenons à l'analyse d'une trame sonore dont le scalogramme est donné sur la figure 58. Remarquons que l'échelle verticale, celle de l'échelle de la décomposition en ondelettes, est aussi **l'axe des "canaux"** puisque à t fixé on a le résultat des filtres à différentes échelles. Or, nous avons vu que ce scalogamme varie pour un même mot prononcé, selon le locuteur par les différences d'attaque des phonèmes, par la différence du rythme etc. **Donc, si on veut extraire des features permettant de discerner deux locuteurs, il faut réaliser une transformation qui relie à t fixé les coefficients le long de l'axe des échelles/l'axe des canaux.** Ceci a été réalisé par exemple par une analyse temps-fréquence dans l'article de J. Anden, V. Lostanlen et S. Mallat [80]. Le signal $x(t)$ est tout d'abord décomposé avec des filtres en odelettes (de type cochléaire) pour donner les scalogrammes des coefficients $|x \overset{t}{*} \psi_\lambda|$ où la convolution est selon l'axe du temps (cf. comme d'habitude), et on va interpréter ce scalogramme comme une image dont on va faire la décomposition en ondelettes bi-dimensionnelles donc permettant de filtrer dans les 2 directions t et $\log \lambda$. En bout de

80. https://www.di.ens.fr/~mallat/papiers/IEEESignalAndenLostanlen.pdf

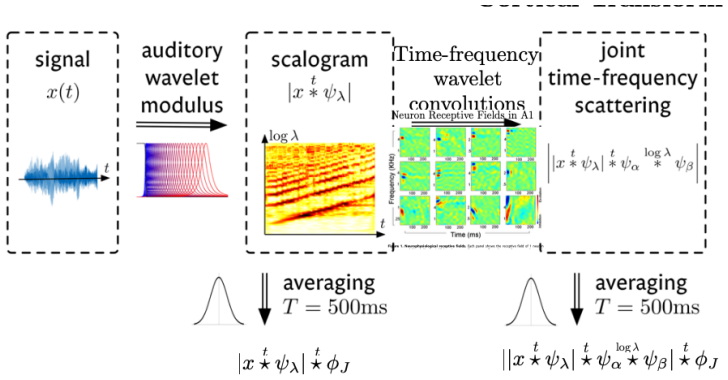

FIGURE 59 – Exemple de traitement d'une décomposition d'un signal 1D par ondeletets, dans lequel le scalogramme est traité comme une image par des ondelettes 2D filtrant ainsi le long de l'axe du temps et de l'axe des échelles qui est celui des canaux.

traitement on dispose non seulement des invariants du signal par moyennage [81] au niveau du scalogramme traditionnel $|x \overset{t}{*} \psi_\lambda| \overset{t}{*} \phi_J$ mais aussi au niveau de l'analyse de l'image temps-fréquence qui comporte une convolution temporelle et "fréquentielle" $||x \overset{t}{*} \psi_\lambda| \overset{t}{*} \psi_\alpha \overset{\log\lambda}{*} \psi_\beta| \overset{t}{*} \phi_J$. Ce qui est remarquable c'est que ce type d'analyse est d'abord apparue en neurophysiologie, dans ce qu'appelle S. Shamma *la transformée corticale* qui réalise un filtrage bi-dimensionnel, identifiée dans le cortex auditif de furets.

S. Mallat présentent d'autres travaux des années 2016-17 sur la reconnaissance de phrase instrumentales (voir MeddleyDB avec 10,000 exemples d'entrainement pour 8 classes: clarinette, guitare électrique, chanteuse, violon, etc) ou de sons urbains (voir UrbanSound8k: 8,000 échantillons pour 10 classes: klaxons, aboiements, souflerie, sirène, etc) qui montrent qu'avec de petites bases d'entrainement, **l'analyse des séries temporelles avec des réseaux de scattering des scalogrammes temps-fréquences**, on obtient des taux d'erreur de mauvaises classifications bien meilleurs (cf. environ 20%) que ceux obtenues soit avec des descripteurs MFCC (état de l'art ère pré-réseaux de neurones) qui obtiennent 40%, soit avec des réseaux convolutionnels ou encore avec un simple réseau de scattering (cf. sans la connexion entre canaux) qui obtiennent environ 30%. Il est clair que se sont des exemples qui datent un peu maintenant mais qui montraient en quoi la corrélation des "canaux" est importante.

81. ϕ_J est une moyenne gaussienne sur 500 ms.

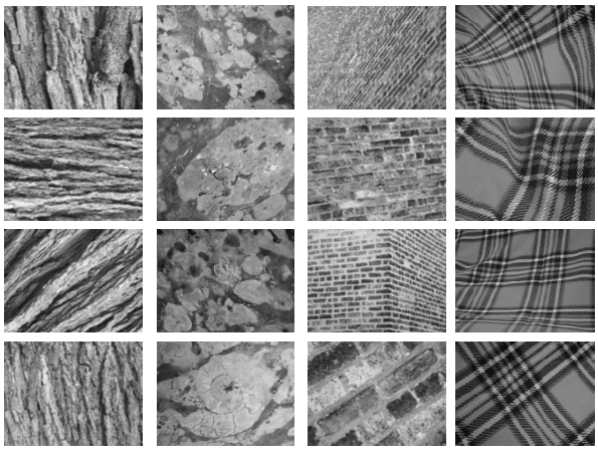

FIGURE 60 – Exemple de textures de la base UIUC.

9.6.4 La classification des textures avec des rotations/zooms

Un dernier exemple [82] concerne la reconnaissance de textures pour lesquelles **les rotations et les variations d'échelles** (cf. zoom) peuvent être de magnitudes très importantes comme on peut l'apprécier sur la figure 60. On débute avec des images, et il faut mettre en œuvre tout d'abord une décomposition en ondelettes ayant les éléments qui permettent de cerner des "bords" dans toutes les directions comme pour la figure 56: on utilise les ondelettes $\psi_{2^j,\theta}$ comme Eq. 178. Ensuite, si l'on veut une invariance par rotation, il faut connecter les résultats le long de l'axe des θ. Si l'on veut une invariance selon le zoom/dézoom (dilatation/contraction) on fait de même sur l'axe des échelles, et on peut combiner les deux (voir figure 61).

Il s'agit donc de capturer la variabilité le long des angles donc concevoir des invariants par rotations tout en étant capables de retenir les détails des structures le long des angles. L'image $x(u)$ (u l'indice du pixel) passée à travers la décomposition en odelettes donne des coefficients $x_j(u,\theta) = |x * \psi_{2^j,\theta}|$, c'est une sortie du premier étage du réseau (W_1) qui correspond à l'analyse passe-bande, l'autre sortie est le résultat du filtre passe-bas qui donne accès à la moyenne des valeurs de pixels $\int x(u)du$. Maintenant, à une échelle 2^j donnée, la collection $x_j(u,\theta)$ peut être vue comme une image 2D indicée par u et θ (cf. on prend des valeurs discrètes d'angles, cf. Eq. 142).

82. Thèse de Laurent Sifre (2014) https://www.di.ens.fr/data/publications/papers/phd_sifre.pdf

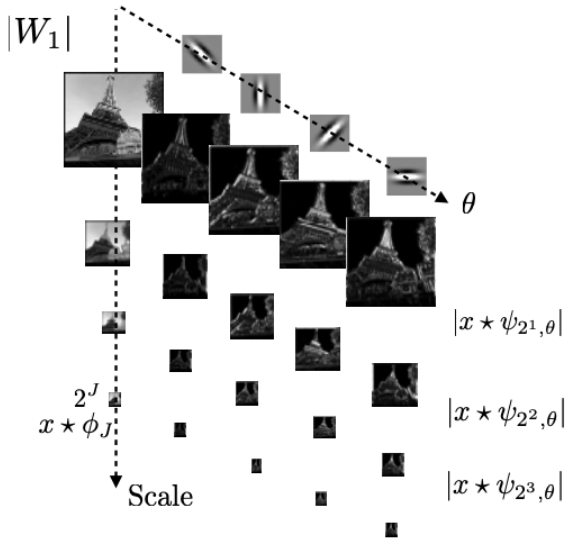

FIGURE 61 – Décomposition en ondelettes de type $\psi_{2^j,\theta}$ d'une image qui permet de mettre en évidence 2 types de canaux que l'on peut filtrer pour obtenir des invariants soit par rotation, soit par dilatation/contraction ou zoom/dézoom, soit les deux à la fois.

Notons que si on applique une rotation d'angle α à l'image d'origine et une translation de c pixels alors les coefficients $x_j(u,\theta)$ deviennent

$$x(u) \to x(r_\alpha(u-c)) \Rightarrow x_j(u,\theta) \to x_j(r_\alpha(u-c),\theta-\alpha) \tag{186}$$

Donc, on constate que les coefficients se déplacent sur l'axe des θ. Pour s'affranchir de cette variation, comment construire un invariant par rotation? Il faut procéder à une convolution qui agit non seulement dans l'espace (u_1,u_2) (cf. indices des pixels en 2D) mais aussi selon l'axe des rotations:

$$x_j \circledast \psi_{j',k}(u,\theta) = \int_0^{2\pi} \left(\iint_{\mathbb{R}^2} x_j(u',\theta') \psi_{2^{j'},\theta}(r_{-\theta'}(u-u'))du' \right) \psi_{2^k}(\theta-\theta')d\theta' \tag{187}$$

On utilise les ondelettes 2D $\psi_{2^j,\theta}$ tournées et translatées dans l'espace des pixels, et une ondelette 1D ψ_{2^k} translatée dans l'axe des θ. Ainsi, on a un schéma de cascade de filtrage comme pour l'audio qui dans le cas de l'image permet d'obtenir des invariants par rotations globales (voir la figure 62). Si maintenant on s'intéresse à l'erreur de classification, on

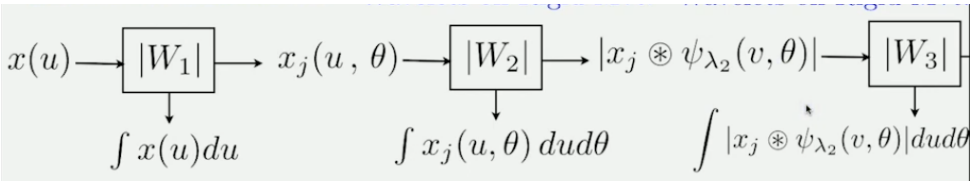

FIGURE 62 – Processing d'une image par une succession de filtres qui dés la seconde étape met en jeu des convolutions selon l'axe des angles pour obtenir des invariants par rotation.

s'aperçoit que si on utilise un réseau de scattering sans prendre en compte la corrélation selon l'axe des angles on obtient 20% d'erreur, alors que si on utilise les invariants par rotations alors le taux tombe à 0.6% ce qui montre l'importance de ces invariants et donc de l'usage des canaux.

9.6.5 Exemple en Chimie Quantique

La prise en compte de ces mouvements 'rigides' est extrêmement importante car ils sont présents partout notamment en Physique et Chimie Quantique. Le but est de calculer l'énergie $f(x)$ d'une molécule à partir de la configuration x, cf. la position et charge de ses constituants. On veut savoir finalement si la molécule est stable ou pas. L'idée de base est qu'au lieu de résoudre les équations de Schrödinger (voir la technique de l'électron unique de Kohn-Sham) pour obtenir la répartition de la densité électronique, on va utiliser des réseaux de neurones profonds tout en incorporant toute l'information *a priori* que l'on connait du problème, pour limiter au mieux le nombre de paramètres "inconnus" qu'il faut obtenir par apprentissage. Or, manifestement, $f(x)$ est invariante par translation et rotation, et il y a des forces à plusieurs échelles caractéristiques (Van der Waals, liaison covalente, etc). Donc, on s'aperçoit que les symétries du problèmes de Chimie ne sont pas très différentes de celles rencontrées pour la classification des textures de la section précédente.

Quand on analyse le problème physique, le modèle de Kohn-Sham écrit l'énergie $E(\rho)$ de la molécule comme une somme de différents contributions dépendantes de la densité électronique ρ:

— un terme d'énergie cinétique,

— un terme d'interaction attractif des électrons par un potentiel effectif,
— un terme d'interaction répulsive d'origine coulombienne entre les électrons,
— et un terme quantique très compliqué à calculer qui va être responsable de la cohésion de la molécule.

et à l'équilibre $f(x)$ est donnée par le minimum de $E(\rho(x))$ où ρ est fonction de la configuration (positions, etc) x des charges en présence.

Maintenant, au lieu de résoudre les équations différentielles du système de Kohn-Sham, on va naïvement essayer d'utiliser les techniques d'apprentissage statistique en tentant de trouver une approximation $\tilde{\rho}(x)$ de $\rho(x)$ et en écrivant une décomposition de $E(\rho(x))$ selon

$$E(\rho(x)) \approx \sum_k w_k \phi_k(\tilde{\rho}(x)) \tag{188}$$

mais on va y mettre **notre connaissance de la physique du problème à travers les descripteurs** ϕ_k. Les poids représentent en quelque sorte des potentiels chimiques que l'on va apprendre à adapter au problème posé à partir d'une base de données de cas traités par résolutions des équations différentielles.

Donc, la première chose à décider est: quel type de descripteur ϕ_k va-t'on utiliser? Nos guides ici sont les symétries du problème. Dans un premier temps, on se fixe une représentation très naïve de la densité électronique, en utilisant la position r_k de chaque charge z_k à la façon d'une somme de fonctions de Dirac centrées sur les atomes:

$$x(u) = \tilde{\rho}(u) = \sum_{k=1}^{d} z_k \delta(u - r_k) \tag{189}$$

Si maintenant, on procède à une décomposition par une ondelette, la convolution par un Dirac redonne l'ondelette donc

$$\rho\left(x * \psi_{2^j,\ell}(u)\right) = \rho\left(\sum_{k=1}^{d} z_k \psi_{2^j,\ell}(u - r_k)\right) \tag{190}$$

(nb. ℓ est un indice d'orientation). Chaque charge k émet en quelque sorte une onde $\psi_{2^j,\ell}(u - r_k)$, et toutes ces ondes interfèrent pour donner des figures d'interférences de plus en plus complexes quand l'échelle devient de plus en plus grande comme illustré sur la figure 63. Ensuite, ces figures d'interférences sont les ingrédients utilisés dans la décomposition de l'énergie du système. Naturellement, comme on a vu dans les paragraphes

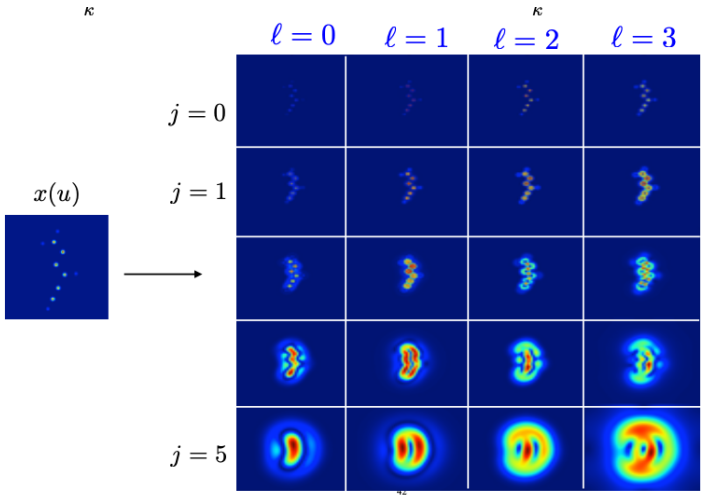

FIGURE 63 – Résultat de $\rho\left(x * \psi_{2^j,\ell}(u)\right)$ (Eq. 190) à partir d'une distribution ponctuelle des charges sur la position des atomes d'une molécule.

précédent, on ne s'arrête pas là car on peut cascader les niveaux de filtrage et calculer au second niveau

$$\rho\left(\rho\left(x * \psi_{2^j,\ell}\right) * \psi_{2^{j'},\ell'}\right) \tag{191}$$

puis on moyenne à une certaine échelle pour obtenir des invariants par translation et rotation, et par une régression linéaire (la seule partir apprise) on essaye de prédire l'énergie.

Les bases de données de chimie quantique (QM9 [83]) contiennent de l'ordre de 130,000 molécules organiques dont on a calculé les configurations et l'énergie par les méthodes traditionnelles. Une fois l'apprentissage fait, on obtient des erreurs avec les réseaux de scattering de l'ordre de 0.5kcal/mol soit du même ordre de grandeur des réseaux convolutionnels où l'on apprend tout (les descripteurs et la régression linéaire). Notons que 0.5 kcal/mol c'est petit et du même ordre de grandeur des erreurs des techniques traditionnelles de résolution des équations de Schrödinger. Mais la question est Pourquoi est-ce si bon? **Où sont passé les aspects quantiques?** Cependant relativisons, ces bases de données de type QM9 sont constituées de molécules à petits nombres d'atomes (cf. max 29 atomes dont 9 lourds), donc on peut dire que le problème résolu ici est du type en traitement d'images de celui de MNIST qui est considéré à présent comme un jeu d'enfant.

83. http://quantum-machine.org/datasets/

FIGURE 64 – Exemples d'images de la base CIFAR-10.

9.7 Échec des réseaux de scattering

Dans les sections précédentes nous avons présenté des cas de figure où l'on peut injecter de l'information *a priori*, et trouver des descripteurs à base d'ondelettes dont les performances sont aussi bonnes que celles des réseaux convolutionnels 'classiques' où l'on ne fait aucun *a priori*[84]. Cependant, il y a des cas où les réseaux de scattering sont moins performants.

Prenons le cas de la base de données CIFAR-10 de 50,000 images 32×32 pixels étiquetées sur 10 classes pour lesquelles il y a une grande variabilité au sein d'une même classe qui n'a rien à voir avec des translations, rotations ou déformations (Fig. 64). Les résultats montrent que l'approche des réseaux de scattering atteint ses limites et c'est le domaine de la recherche actuelle. S. Mallat relate les travaux d'Edouard Oyallon [85] dont le séminaire a été annulée pour cause de COVID-19. Si on prend un réseau de neurones 'classique' tout à fait commun, on obtient un taux d'erreur commun de 7%; par contre si on utilise un réseau de scattering avec la partie filtres en ondelettes intégrant tous les groupes de symétries *a priori* qui n'a pas d'apprentissage et une partie de régression qui elle est apprise, alors on plafonne à 20% d'erreur.

Si on fait un bilan, là où les réseaux de scattering fonctionnent bien sont les exemples où les groupes de symétries que l'on connait décrivent bien la variabilité du problème: digits (10 classes), texture (60 classes), chimie quantique à petits nombres d'atomes. Par contre, dés que l'on s'attaque à des problèmes plus complexes comme la classification d'ImageNet (environ 2M d'images 256×256 pixels, 1000 classes) alors: en 2012 AlexNet qui a été le coup de tonnerre de la discipline avait une erreur de 16.4% (maintenant on a des réseaux convolutionnels qui ont des taux d'erreur de l'ordre de 1% inférieur au taux

84. Attention: l'architecture est un *a priori* à garder en tête.
85. https://edouardoyallon.github.io/thesis.pdf

d'erreur humain [86]), et les réseaux de scattering plafonnent à 60% c'est-à-dire bien moins bon que l'état de l'art avant AlexNet qui eux avaient un taux de 26% (voir Cours 2019). La question est alors: q**u'est-ce qui est appris par les réseaux convolutionnels et qui fait défaut aux réseaux de scattering?**

Il y a une façon d'aborder cette question, en ce souvenant qu'historiquement ce domaine était appelé "pattern recognition" ou "reconnaissance de forme". C'est-à-dire, que l'on s'attend malgré tout à devoir **reconnaitre des structures** [87] que l'on aurait besoin d'apprendre pour résoudre les problèmes. L'approche mathématique consiste à utiliser la notion de **dictionnaires** et d'en trouver qui réalisent une **description parcimonieuse** du problème. Typiquement, on veut projeter x sur un dictionnaire $\mathcal{D}$ constitué de k patterns $\{D_k\}_{k\leq d}$ de telle façon que

$$x = \sum_{k=1}^{d} D_k z_k = \mathbf{D}.\mathbf{z} \tag{192}$$

et la représentation est parcimonieuse si $(z_k)_{k\leq d}$ est constitué essentiellement d'éléments nuls. Pour obtenir une telle parcimonie (sparsity) on va utiliser la technique de la **régularisation par la norme L1** pour contraindre l'espace dans lequel on veut trouver z tel que $\mathbf{D}.\mathbf{z}$ approxime x (voir Cours 2018 et 2019), c'est-à-dire que

$$\tilde{z} = \underset{z}{\operatorname{argmin}} \ \|x - \mathbf{D}.\mathbf{z}\|_2^2 + \alpha\|\mathbf{z}\|_1 \tag{193}$$

C'est un problème d'optimisation convexe que l'on peut résoudre avec des algorithmes dont la convergence est garantie. Il se trouve que certains de ces algorithmes sont itératifs et peuvent être implémentés sous forme de réseau de neurones, ils apparaissent alors comme une cascade d'opérateurs constitués d'un seul dictionnaire, suivis d'un rectificateur (non-linéarité) comme sur le schéma [88] de la figure 65.

Si la machinerie est en place, la question en suspend est alors de déterminer **les patterns** qui vont être utiles à la résolution du problème? On ne les connait pas à l'avance et **on a besoin de les apprendre**. L'idée est alors de construire une grande matrice qui représente le dictionnaire et de faire une optimisation par descente de gradient en utilisant un lot

86. Avec les bémols à cette comparaison.
87. voir la grammaire de Chomsky du Cours de 2019
88. Voir par exemple les travaux de John Zarka, Louis Thiry, Tomás Angle et S. Mallat https://www.di.ens.fr/~mallat/papiers/Zarka-ICLR2020.pdf.

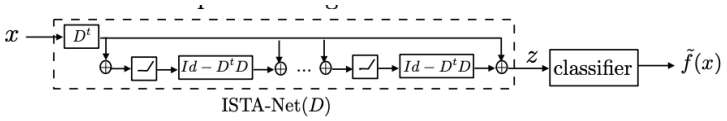

FIGURE 65 – Exemple d'algorithme (ISTA-Net) de décomposition de x en représentation parcimonieuse avec une cascade d'opérateurs constitués d'un dictionnaire D et suivis de rectificateur pour garantir la positivité de z.

d'entrainement $\{x_i, y_i\}$ telle que l'on minimise l'erreur (ℓ pour *loss*)

$$\ell(D) = \sum_i \ell(y_i, \tilde{f}_D(x_i)) \tag{194}$$

Le résultat est bien une matrice des structures élémentaires permettant de décrire d'une manière parcimonieuse la variable x.

Une illustration est donnée par une comparaison des résultats obtenus sur la base d'ImageNet (Tab. 2):

1) tout d'abord on peut mettre en œuvre un réseau de scattering où on essaye d'incorporer toute l'information *a priori* que l'on connait sur les symétries du problème et on optimise un classificateur qui vient donner la réponse finale du réseau. Le résultat est de 60% d'erreur.
2) on peut également utiliser la décomposition de type ISTA-Net où on apprend conjointement primo les dictionnaires pour obtenir une description parcimonieuse, et secundo les paramètres du classificateur. Le résultat est du même ordre, cf. 50%, en gros on a rien gagné.
3) enfin on peut combiner les deux réseaux, à savoir que dans une première étape on utilise la connaissance *a priori* dans un réseau de scattering (1), mais **cette fois les invariants qui en sortent sont passés à travers un réseau de dictionnaires** (2) dont la sortie est utilisée pour faire la classification. L'optimisation conjointe du dictionnaire et du classificateur, donne finalement **un taux d'erreur de 18%** comparable au score du réseau AlexNet de 2012.

Comment ce passage de 50-60% à 18% peut-il s'interpréter? Si on reprend le schéma (2), on constitue des dictionnaires directement à partir des images d'ImageNet, mais alors face à la grande variabilité au sein même de chaque classe, le nombre de patterns

Modèle	Apprentissage	taux d'erreur
Scattering Net (1)	non	60%
Dictionnaires (2)	oui	50%
Scattering + Dictionnaire (3)	mixte	18%

TABLE 1 – Comparaison de différentes types de réseaux (voir texte) sur la classification d'ImageNet. Pour mémoire le réseau AlexNet de 2012 avait un taux d'erreur du même ordre que le (3).

du dictionnaire est gigantesque, et les apprendre efficacement demanderait un nombre colossal d'échantillons par classe. **Si par contre, on linéarise par rapport aux symétries que l'on connait dont les difféomorphismes, on opère une réduction très importante de la variabilité, ce qui permet une recherche efficace des patterns qui doivent capturer les symétries restantes**. Semble-t'il la variabilité à suffisamment décrue pour que une telle recherche de patterns soit effective et fasse tomber le taux d'erreur d'un facteur 3.

L'avantage de l'architecture (3) est que l'on sait interpréter les différentes couches du réseau, de plus on a qu'une seule matrice à apprendre. Par contraste, si les architectures à 300 couches de type ResNet ont des taux d'erreur d'environ 3%, celles-ci restent pour le moment difficilement interprétables. Cependant, il faut dire que l'on ne comprend pas totalement la réduction d'un facteur 3 de l'erreur opérée entre les architectures de type (1)-(2) et la combinaison (3). Quelle est la nature mathématique qui opère? il y a des hypothèses mais pas de bon modèle mathématique qui explique bien le phénomène.

9.8 Conclusion du cours 2020

S. Mallat termine le cours de cette année en formulant quelques observations. Tout d'abord nous avons vu que les réseaux de neurones profonds convolutionnels (CNN) obtiennent des résultats spectaculaires en très grande dimension et cela pour des problèmes de types très différents (imagerie, son/voix, chimie quantique, langage, etc). C'est d'autant plus spectaculaire qu'*a priori*, si la fonction à approximer n'a pas une régularité forte ce problème est difficile voire impossible. Incontestablement, ces architectures CNN sont capables d'apprendre quelque chose sur les objets traités en particulier sur les symétries

du problème. Il reste cependant un mystère sur le **Pourquoi** d'une telle efficacité, et en particulier Pourquoi une même architecture est capable de capturer la régularité de fonctions très complexes et diverses. Pourquoi ces problèmes partagent-ils le même type de régularité/symétrie que capture les CNN?

Cette année, nous avons tenter de montrer qu'il y a trois approches pour expliquer la régularité de la fonction sous-jacente que l'on tente d'approximer.

1) Le premier axe s'attache à la **séparabilité**. Par la séparation d'échelles à l'aide d'une analyse multi-résolutions, on obtient des structures certes mais surtout on peut linéariser par l'action de **groupes de symétries** (ex. translation, rotation, déformation) et obtenir des **invariants** qui sont des structures très rigides et diminuent la variabilité du problème (réduction de dimension). C'est une attitude très courante en Physique de Particules par exemple. Cependant, quand on arrive à l'échelle macroscopique, on rencontre la notion de **patterns** presque inévitablement. On le sent bien en Chimie par exemple, où les briques élémentaires des interactions de bases fussent-elles quantiques fonctionnent bien à l'échelle microscopique par exemple pour expliquer l'apparition de bandes de conduction ou autres. Mais, quand on s'intéresse des propriétés de molécules complexes, on constate que la Chimie 'traditionnelle' non-quantique rend malgré tout compte de propriétés génériques que l'on peut qualifier de "patterns" que l'on a appris "empiriquement". Le même type de raisonnement vaut pour l'analyse de sons où l'oreille reconnait des structures qui ne sont pas au niveau d'une seule note éparse, mais pour comprendre des mélodies il faut une phase d'apprentissage. Idem pour l'analyse d'images où l'on reconnait des traits communs à tel ou tel visages assez rapidement par exemple, mais pour reconnaitre telle ou telle essence d'arbres il faut une phase d'apprentissage. Ce qui vaut pour l'éducation de l'ouïe, l'œil, vaut certainement pour d'autres domaines [89], ce qui aboutit à la notion de **dictionnaire** dont les éléments sont appris, et de **représentation parcimonieuse**.
2) Le second axe qui a été développé dans le cours et les séminaires, c'est le lien avec la **neurophysiologie**. On l'a vu pour l'appareil auditif et visuel, et il est très spectaculaire de constater les similitudes d'architectures (informatique/neurophysiologique) qui sont mises au jour par des études indépendantes ou mixtes entre les deux communautés. Cependant, il faut faire attention à ne pas sur-jouer la similitude: les

89. NDJE: On pourrait d'ailleurs en dire autant pour l'apprentissage des concepts en mathématiques.

neurones artificiels n'ont pas le même fonctionnement d'un neurone biologique, le cerveau humain est bien plus complexes qu'un réseau de neurones artificiels, etc. Mais les similarités sont suffisamment intéressantes pour motiver de plus amples investigations.

3) Enfin, il y a un autre thème que nous n'avons pas encore développé durant les cours jusqu'à présent mais que nous développerons dans le futur, c'est le lien avec la **Physique**. Par nature, la Physique jusqu'à récemment était la seule **science de la grande dimension**. En particulier, la Physique Statistique traite de problème où le nombre d'Avogadro est l'étalon. C'est le domaine où des modèles mathématiques ont été développés pour comprendre des systèmes dynamiques de grande dimension. On voit des interfaces se développer récemment sur des problèmes où la Physique a du mal: on a évoqué la Chimie Quantique, il y a aussi par exemple la compréhension de la turbulence en dynamique des fluides qui restent un problème encore ouvert depuis les travaux de Kolmogorov des années 40. Du coté des réseaux de neurones, il semble apparaître des possibilités de simuler de la turbulence qui motive des recherches communes.

Pour finir, l'étude des réseaux de neurones du coté mathématiques reste un problème essentiellement ouvert. On n'a pas vraiment d'échelle de complexité qui permettrait par exemple de concevoir une architecture (nombre de couches, de quelle type, nombre de neurones par couche, etc) connaissant le problème à résoudre en étant certain quelle satisfasse à la question posée. On aimerait disposer d'outils mathématiques pour capturer la régularité/symétrie des fonctions en très grande dimension. Enfin, on aimerait avoir des théorèmes d'approximation qui garantissent le niveau de l'erreur commise et la stabilité des résultats (cf. exemples adversaires). S. Mallat nous pointe un article [90] dans lequel il passe en revue des notions élaborées dans le cours.

L'an prochain le thème envisagé est celui de la Parcimonie.

90. https://www.di.ens.fr/~mallat/papiers/RSTA2015Published.pdf