Module de BCI :
Outils et applications pour le signal, les images et le son
OASIS (SI101)
2016-2017

Saïd Ladjal (said.ladjal@telecom-paristech.fr)
et Pascal Bianchi (pascal.bianchi@telecom-paristech.fr)

Table des matières

Présentation du polycopié

Ce polycopié regroupe l'essentiel de ce qui sera vu en cours d'OASIS plus quelques compléments.

Après une brève introduction aux signaux (qui ne sont rien d'autre que des fonctions), le premier chapitre introduit et étudie la notion de système linéaire invariant. Il s'agit d'une modélisation naturelle de beaucoup de processus physiques liant un signal de sortie à un signal en entrée. Dans ce chapitre, on verra qu'avec de très faibles hypothèses sur le système qui lie la sortie à l'entrée, on arrive à déduire une caractérisation extrêmement puissante de tous les systèmes linéaires et invariants.

À la fin du premier chapitre, nous verrons que les systèmes linéaires et invariants (SLI, dans la suite) ont un comportement privilégié vis-à-vis d'un certain type de fonctions que nous appellerons "ondes pures" ou "ondes de Fourier" (ces ondes sont des vecteurs propres des SLI). Ceci nous amènera à définir la transformation de Fourier qui est, très grossièrement, la décomposition sur la "base" des ondes de Fourier d'un signal. Nous verrons en particulier les rapports entre SLI et Fourier. L'étude des propriétés fondamentales de transformée de Fourier constituera le chapitre 2.

Ensuite, on introduit l'outil transformée en Z qui est une réinterprétation de la transformée de Fourier à temps discret. Elle permet, essentiellement, l'analyse d'une classe de filtres très utile : les filtres récursifs.

Au quatrième chapitre, nous abordons le problème de l'échantillonnage. La question posée et à laquelle nous tentons de répondre est "Comment retrouver un signal défini sur $\mathbb{R}$ à partir de la connaissance de ses valeurs sur $\mathbb{Z}$?" Bien que ce problème semble insoluble, nous verrons qu'avec les bonnes hypothèses sur le signal nous arriverons à le résoudre en nous basant sur ce qui a été vu aux précédents chapitres. Ce chapitre est extrêmement important du fait que toutes les informations multimédias, de nos jours, sont stockées sur un support numérique et donc discret.

Les chapitres 5, 6 et l'annexe B portent sur les outils de la compression des signaux naturels en mettant l'accent sur la compression d'images. On introduit une transformée proche de la transformée de Fourier et qui est adaptée aux signaux finis non périodiques (chapitre 5). Puis on introduit les bases de la compression des signaux numériques. La quantification est en annexe.

Enfin, le dernier chapitre introduit quelques notions sur les processus aléatoires discrets. Un processus aléatoire discret est une fonction définie sur $\mathbb{Z}$ à valeur dans l'espace des variables aléatoires sur un espace probabilisé. On ajoute à cette définition quelques hypothèses ou contraintes (par exemple : que toutes les variables aléatoires du processus soient indépendantes et identiquement distribuées). Les processus modélisent tout ce qui, dans une communication ou acquisition de signal, n'est pas facilement quantifiable ou encore ce qui est par nature aléatoire. On verra comment les outils définis peuvent nous

aider à résoudre du “mieux possible” un problème de débruitage ou encore à synthétiser une voix humaine.

Bonne lecture.

Chapitre 1

Les systèmes linéaires et invariants (SLI)

Dans ce chapitre, on étudie les systèmes linéaires invariants (SLI dans la suite). Nous commençons par les définir d'un point de vue mathématique, puis nous donnons quelques exemples de tels systèmes. Les exemples seront de deux types. Les exemples purement mathématiques et des exemples issus de l'étude d'un système physique réel (et la modélisation qui en découle). Ensuite, nous passons à la caractérisation de ces systèmes par le biais d'un outil mathématique appelé convolution. Enfin, nous verrons que tous les SLI ont un comportement particulier vis-à-vis des ondes de Fourier (que nous définirons ici en prévision du prochain chapitre).

1.1 Les signaux, exemples et définition

Une définition naïve des signaux serait : tout type de phénomène physique perceptible par nos sens visuel et auditif. Par exemple, un son est caractérisé par la variation dans le temps de la pression de l'air en un point de l'espace. Ce son est perceptible par nos oreilles. Une image est une variation dans un espace bidimensionnel de la luminosité. Il est perceptible par nos yeux.

Quelques exemples plus précis :

1.1.1 Signaux sonores

Un signal sonore est la mesure de l'évolution dans le temps de la surpression de l'air en un point donné de l'espace. Le son peut donc être vu comme une fonction du temps continu, c'est-à-dire une fonction de $\mathbb{R}$. La figure 1.1 montre un exemple de signal sonore. On remarque, par exemple, que ce signal semble être symétrique par rapport à l'horizontal. C'est une caractéristique des signaux sonores d'être de somme nulle (sur un temps assez long). Sur cette figure on remarque que l'échelle du temps est trop grossière et ne permet pas de percevoir les variations fines du signal. Sur la figure 1.2 nous avons zoomé le même signal sur une petite période de temps (de l'ordre 0,1 seconde). À cette échelle-ci, on peut voir les variations du signal. On constate même que le son est (sur cette courte durée) pseudo-périodique avec une fréquence de 100Hz (1/0.01s). Dans la suite du

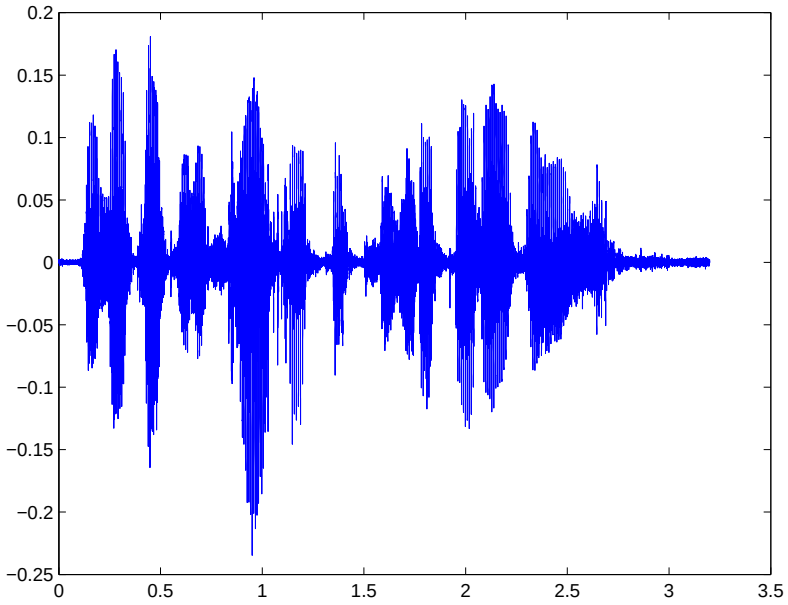


FIGURE 1.1 – **Un exemple de signal sonore sur quelques secondes. Il est dit "Il était une fois un petit chaperon rouge". On remarque un certain nombre de pauses entre les mots. Affiché à cette échelle, le graphique est peu lisible.**

chapitre, nous verrons comment le signal sonore généré par une source est transformé, par l'environnement, en un autre signal sonore qui sera, par exemple, enregistré.

1.1.2 Le Compact Disc

Un compact disc musical contient une succession de valeurs inscrites de manière numérique sur le support. Pour chaque canal stéréo, le compact disc contient 44100 valeurs par seconde de musique. Il s'agit de l'échantillonnage d'un signal sonore. On peut dire que le signal contenu dans un CD est une fonction d'un temps discret, on peut indexer les échantillons par l'ensemble des entiers relatifs $\mathbb{Z}$.

La figure 1.3 montre un extrait de 100 échantillons d'un CD musical. Le passage d'un signal sonore indexé par un temps continu à une suite d'échantillons indexée par les entiers est étudié dans le chapitre intitulé "Échantillonnage". Nous y expliquerons aussi la raison de ce choix de 44100 échantillons par seconde.

1.1.3 Le disque vinyle

Un disque vinyle est un support pour la musique sur lequel[1] le signal sonore est enregistré de la manière suivante : une spirale est gravée sur une galette de vinyle. La spirale parfaite représenterait un son parfaitement nul. Pour stocker un son non nul, la spirale est légèrement modifiée (perturbée) pour refléter le signal sonore.

1. Nous parlons du vinyle mono, plus simple que le vinyle stéréo

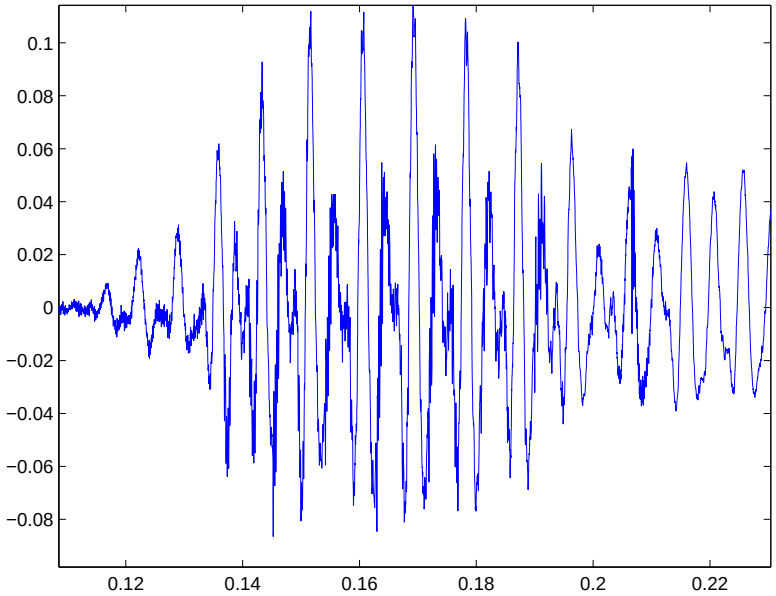


FIGURE 1.2 – **Nous avons zoomé sur une partie du signal précédent d'une durée d'environ 0,1 seconde. Remarquer comme le signal est caractérisé par une pseudo-périodicité. À cette échelle, on peut presque mesurer la pseudo-période qui vaut à peu près 0,01seconde, soit une féquence de 100Hz. Le chapitre "échantillonnage" élucidera la bonne échelle à laquelle il faut analyser un son audible.**

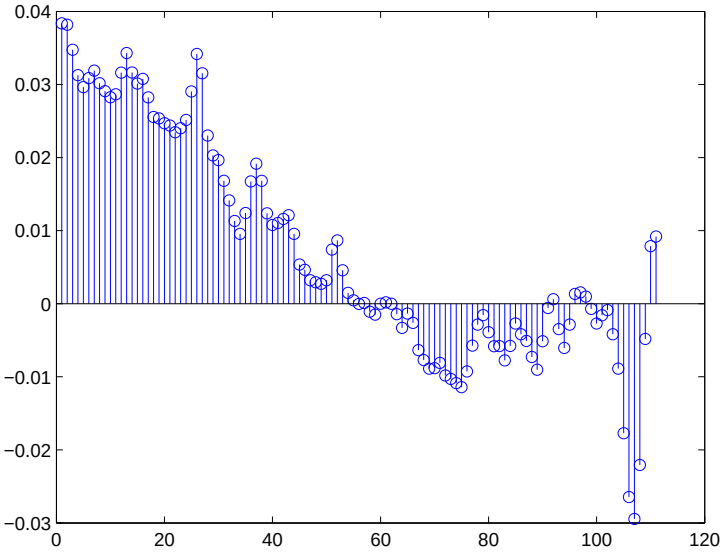


FIGURE 1.3 – **Une suite de 110 échantillons d'un CD. Le temps est devenu discret. Cet extrait correspond à (110/44100) secondes de musique. L'échelle à laquelle sont pris les échantillons semble être suffisante pour capturer les variations du son enregistré sur CD.**

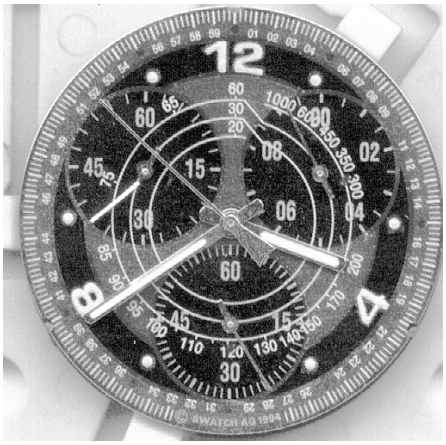


FIGURE 1.4 – **Photographie d'une montre. Un point est d'autant plus blanc qu'il a reçu plus de lumière. Un point qui ne reçoit pas de lumière est noir. Dans le chapitre "Image" nous verrons comment les images sont acquises, et comment leur processus d'acquisition peut être vu, approximativement, comme un SLI.**

Lorsque l'aiguille passe dans le sillon, le mouvement que lui impose la spirale parfaite est négligeable, par contre les petites perturbations qui sont bien plus brutales que la spirale parfaite causent un mouvement de l'aiguille qui est mesuré puis amplifié avant d'arriver aux enceintes.

1.1.4 Les images

Les images sont le signal auquel nos yeux sont sensibles. On peut les regarder comme une fonction à deux variables et à valeurs réelles [2]. Cette valeur est d'autant plus grande que le point est lumineux (blanc dans une image en niveaux de gris).

Les figures 1.4 et 1.5 montrent une photographie et sa représentation en 3D en tant que fonction d'élévation à deux variables.

1.1.5 Distribution de température

Soit une barre droite de dimension infinie. La donnée de la température en chaque point s'appelle distribution de température. Cette distribution de température évolue dans le temps par diffusion. Nous reviendrons sur cet exemple pour présenter un cas de SLI.

1.1.6 Définitions

Dans les exemples que l'on a énoncés, nous avons en fait rencontré des fonctions définies sur des ensembles divers. La définition (mathématique) que nous prendrons pour les signaux est assez simple, **un signal est simplement une fonction**.

D'abord, nous listons les ensembles de définition possibles pour les signaux auxquels nous nous intéressons. Pour chacun de ces ensembles de définition nous précisons, si ce n'est pas clair, l'opération d'addition entre deux de leurs membres.

2. Cela est valable pour une image en niveaux de gris. Pour une image couleur, il faut considérer trois fonctions distinctes, une pour le niveau de rouge, une pour le vert et une pour le bleu.

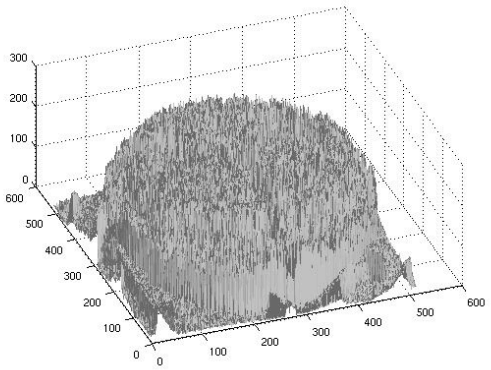


FIGURE 1.5 – **La même photographie qu'à la figure 1.4 représentée comme une surface de l'espace. Les coordonnées x et y sont les indexes de la position sur le plan de la photographie et la coordonnée z (altitude) est la luminosité en ce point. Il est très difficile de reconnaître l'image originale à partir d'une telle représentation, elles sont pourtant équivalentes.**

Définition 1.1. Les ensembles de définition des signaux
La liste des ensembles de définition utilisés pour les signaux, dans ce cours, est la suivante :

1. *L'ensemble des réels,* $\mathbb{R}$*. On définit sur* $\mathbb{R}$ *de manière naturelle un signal sonore, ou un signal électrique, la variable étant vue comme le temps.*
2. *Le plan* $\mathbb{R}^2$*. Dans ce cours, on l'utilisera comme domaine de définition des images. On peut aussi considérer tous les espaces* $\mathbb{R}^n$ *auxquels les résultats génériques de ce cours s'appliquent aussi.*
3. *L'ensemble des entiers relatifs* $\mathbb{Z}$*. Il est utilisé par exemple pour définir une suite très longue d'échantillons sonores sur un CD.*
4. *La grille* $\mathbb{Z}^2$*. On définit dessus les images échantillonnées.*
5. *Le tore* $\mathbb{R}/\mathbb{Z}$ *qui peut se représenter par l'intervalle* $[0, 1[$ *(ou l'intervalle* $[-1/2, 1/2[$ *ou tout autre intervalle de longueur 1). L'opération d'addition est l'addition sur* $\mathbb{R}$ *modulo* 1*. Par exemple, sur cet ensemble,* $0,75+0,3 = 0,05$ *ou encore* $0,5+0,5 = 0$*. Sur cet ensemble se définissent naturellement les signaux périodiques. En effet, une fonction périodique sur* $\mathbb{R}$ *n'a pas besoin d'être connue sur un intervalle plus long que sa propre période pour être connue partout.*
6. *L(es)'ensemble(s) fini(s)* $\{0, \cdots, N-1\}$ *à* N *éléments. que nous noterons aussi* $\mathbb{Z}/N\mathbb{Z}$*. Sur cet ensemble, nous prenons comme opération d'addition l'addition sur* $\mathbb{Z}$ *modulo* N*. Par exemple* $N-1+2 = 1$ *ou encore* $N-1+1 = 0$*. Sur cet ensemble seront définis, par exemple, les échantillons d'un signal discret et périodique.*

Nous explicitons dans la définition qui suit les noms précis des signaux suivant leur domaine de définition. Noter que dans tous les domaines de définition considérés, il existe une opération '+' pour laquelle il existe un élément neutre et tout élément possède un inverse (on parle de structure de groupe).

Définition 1.2. Les signaux
Un signal est une fonction définie sur l'un des espaces listés ci-dessus. Suivant l'ensemble

de définition, on peut trouver, accolé au mot "signal", une précision indiquant son ensemble de définition.

(i) Sur $\mathbb{R}$ on parle de signaux à temps continu.

(ii) Sur $\mathbb{Z}$ on parle de signaux à temps discret.

(iii) Sur $\mathbb{R}^2$ et $\mathbb{Z}^2$, dans ce cours on parlera d'image ou d'image échantillonnée.

(iv) Sur le tore $\mathbb{R}/\mathbb{Z}$ où l'intervalle $[0,1[$ (avec la règle d'addition vue plus haut) on parle de signaux 1-périodiques.

(v) Sur un ensemble $\mathbb{Z}/N\mathbb{Z}$ on parle de signaux discrets finis.

Remarque 1.3. Dans les chapitres suivants (Transformations de Fourier) nous classerons les signaux (fonctions) suivant certains critères d'intégrabilité (sommable, d'énergie finie, bornée). Dans le présent chapitre, on se contente d'une présentation formelle qui s'applique à tous les types de signaux de manière générique. Nous parlons simplement de fonctions.

Avertissement : Dans la suite, nous allons présenter les Systèmes Linéaires et Invariants. Pour une première étude nous les présentons seulement dans le cas de signaux définis sur $\mathbb{Z}$ afin d'alléger les notations et les démonstrations. Vous trouverez dans la suite du chapitre une généralisation des concepts vus pour les suites définies sur $\mathbb{Z}$ à tous les autres types de fonctions que nous étudions dans ce cours.

1.2 Les Systèmes Linéaires et Invariants (SLI) pour les signaux discrets infinis

Lors d'une communication, un émetteur émet un signal qui passe dans un canal de transmission avant d'être reçu par le récepteur. La relation qui lie le signal reçu au signal émis fait l'objet de ce premier cours. Dans beaucoup de cas, le canal a deux caractéristiques principales : il ne varie pas dans le temps et il est linéaire. Ce sont ces cas-là que nous étudions ici. Le fait que le canal ne varie pas dans le temps a pour conséquence que si deux signaux qui ne diffèrent que d'un décalage temporel sont présentés à l'entrée du canal, alors les sorties associées à ces deux signaux ne diffèrent entre elles que du même décalage temporel. La conséquence de la linéarité sera d'obtenir une caractérisation mathématique simple de tous les SLI, la convolution.

Définition 1.4. Translatée d'une suite
Soit u une suite à valeurs complexes définie sur $\mathbb{Z}$ de terme général u_n (évidemment, u peut être vue comme une fonction de $\mathbb{Z}$ vers $\mathbb{C}$ et on pourrait noter $u(n)$, malheureusement cela n'est pas la convention usuelle pour noter les suites...). Soit $m \in \mathbb{Z}$ un entier relatif. On définit la suite v que l'on appelle m-translatée de u par

$$\forall n \in \mathbb{Z},\ v_n = u_{n-m}$$

ou encore, en notation fonctionnelle classique,

$$\forall n \in \mathbb{Z},\ v(n) = u(n-m).$$

Définition 1.5. Espace de suites stable par translation
On dit qu'un sous espace vectoriel V de l'espace vectoriel des suites est stable par translation si

$$\forall u \in V \ \forall m \in \mathbb{Z}, \quad \text{la } m\text{-translatée de } u \text{ est aussi dans } V$$

Autrement dit, si une suite est dans V toutes ses translatées sont aussi dans V.

Définition 1.6. Système linéaire invariant (SLI) pour les suites
Soient V et W des sous-espaces stables par translation de l'espace des suites. Soit T une application de V dans W. On dit que T est un système linéaire et invariant (SLI) si T vérifie les conditions suivantes :

(i) La fonction T est linéaire.

$$\forall u, v \in V, \ T(u+v) = T(u) + T(v)$$

ce qui se traduit sur les termes des suites, en notation fonctionnelle,

$$\forall u, v \in V \ \forall n \in \mathbb{Z}, \ (T(u+v))(n) = (T(u))(n) + (T(v))(n)$$

(ii) La fonction T est invariante par translation, c'est-à-dire que si $u \in V$, m un entier et v la suite m-translatée de u ($v \in V$ par stabilité de V) alors

$$T(v) = m\text{-translatée de } T(u)$$

ou encore, en notation fonctionnelle,

$$\text{si } (\forall n, \ v(n) = u(n-m)) \ \text{ Alors } \ (\forall n \in \mathbb{Z}, \ (T(v))(n) = (T(u))(n-m))$$

Autrement dit si T fait correspondre à la suite u une suite w. Alors T fait correspondre à la suite u décalée de m, la suite w décalée de m. Si l'entrée est décalée, alors la sortie aussi est décalée dans la même proportion.

1.2.1 Exemples

Moyenne glissante
On se donne une série temporelle qui représente la valeur journalière d'une grandeur. Cette grandeur peut être la température en un lieu donné, mesurée à une heure fixe, ou bien la valeur d'un indice boursier. On la note u (u_n est la grandeur au jour numéro n).

La moyenne glissante sur un mois est définie comme la moyenne sur les trente derniers jours de cette grandeur. On la note v_n.

On remarque que v dépend de manière linéaire de u. En effet, si u^1 et u^2 sont deux suites, la moyenne de leur somme est égale à la somme des moyennes. Par ailleurs, si on décale u de m jours alors v est décalée de m jours aussi. La suite v dépend donc de manière invariante par translation de u.

La relation qui donne v à partir de u est donc un SLI.

La formule de calcul de v à partir de u est

$$v_n = \frac{1}{30}\left(u_n + u_{n-1} + u_{n-2} + \cdots + u_{n-29}\right)$$

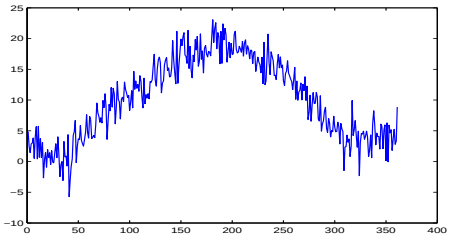


FIGURE 1.6 – **Évolution sur un an des températures.**

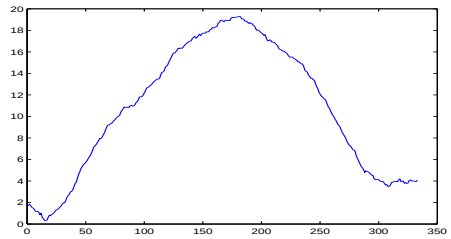


FIGURE 1.7 – **Moyenne glissante sur le mois passé de la température.**

Les figures (1.6 et 1.7) montrent que les évolutions de v sont moins brutales que celles de u. Le filtre moyenneur est un filtre **passe-bas**, c'est-à-dire qu'il favorise les évolutions lentes et gomme les variations rapides.

Variations journalières d'un indice boursier

Si u est une suite de valeurs d'un indice boursier, on peut s'intéresser aux variations d'un jour à l'autre de cet indice. Cette variation journalière est elle aussi une suite et est définie pour chaque jour comme la différence entre la valeur de l'indice au jour considéré moins la valeur au jour d'avant. On la note v.

Cette définition seule permet de déterminer que le lien entre v est u est un SLI. Linéaire, car la différence entre sommes est égale à la somme des différences entre termes et invariante par translation, car le mode de calcul ne change pas d'un jour à l'autre.

La formule qui donne v en fonction de u est

$$v_n = u_n - u_{n-1}$$

Les figures (1.8 et 1.9) montrent que le graphique de v est plus chahuté que celui de u. On dit que u a subi un filtrage **passe-haut**, c'est-à-dire un filtrage qui privilégie les variations rapides du signal et gomme les tendances à grande échelle.

1.2.2 Caractérisation universelle des SLI

Dans cette partie nous allons montrer que tous[3] les SLI (sur $\mathbb{Z}$) s'écrivent sous une forme très simple.

3. Mathématiquement parlant cela n'est pas tout à fait vrai. On peut construire des SLI qui n'obéiront pas à cette règle, mais leur construction fait intervenir l'axiome du choix, ce qui les rend très éloignés du monde physique. Vous pouvez, à ce sujet, consulter l'encyclopédie wikipedia au sujet de "Banach limit"

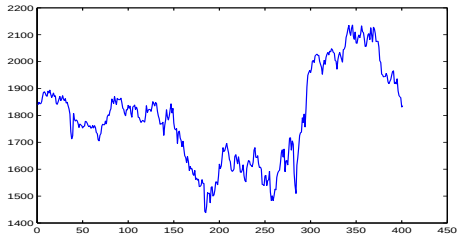


FIGURE 1.8 – **Évolution sur un an d'un indice boursier.**

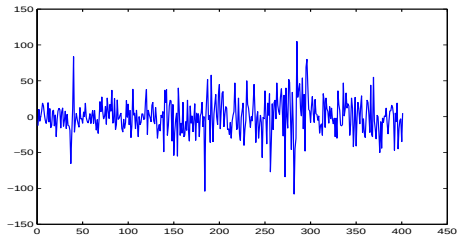


FIGURE 1.9 – **Variations journalières de l'indice boursier.**

Définition 1.7. Suites à support fini
Une suite u est dite à support fini si $u_n = 0$ sauf pour un nombre fini d'entiers relatifs n. En particulier, pour une telle suite

$$\exists N \in \mathbb{N},\ \forall |n| \geqslant N,\ u_n = 0$$

et u tend vers 0 en $+\infty$ et $-\infty$.

L'espace des suites à support fini est clairement un espace vectoriel et il est bien stable par translation.

Proposition 1.8. Caractérisation des SLI pour les suites à support fini

Si T est un SLI qui transforme une suite à support fini en une autre suite à support fini. Alors il existe une suite à support fini, h, telle que (on note $v = T(u)$ la sortie associée à la suite u)

$$\forall n \in \mathbb{Z},\ v_n = \sum_{m\in\mathbb{Z}} u_m h_{n-m} = \sum_{l\in\mathbb{Z}} h_l u_{n-l}$$

On dit que h est la **réponse impulsionnelle** *du SLI T. Notez que les sommes infinies en présence sont en fait finies (car les suites u et h sont à support fini).*

Démonstration.
Soit δ la suite définie par

$$\forall n \in \mathbb{Z}, \delta_n = \begin{cases} 0 & \text{si} \quad n \neq 0 \\ 1 & \text{si} \quad n = 0 \end{cases}$$

Une telle suite est parfois appelée **impulsion en zéro**, ou **suite de Dirac**, elle est bien à support fini. On note h la suite qui lui est associée par le SLI T

$$h = T(\delta)$$

(D'où le nom de réponse impulsionnelle. h est la réponse du SLI à l'impulsion δ).

Pour tout entier relatif m on note δ^m la suite δ translatée de m

$$\forall n \in \mathbb{Z}, \delta_n^m = \delta_{n-m} = \begin{cases} 0 & \text{si} \quad n \neq m \\ 1 & \text{si} \quad n = m \end{cases}$$

(La notation δ_n^m pour signifier 1 si n et m sont égaux ou 0 sinon, est appelée **symbole de Kronecker**)

Si on appelle h^m la suite h translatée de m, l'invariance par translation des SLI nous permet d'écrire

$$\forall m \in \mathbb{Z}, T(\delta^m) = h^m$$

Or, pour toute suite u à support fini on a

$$u = \sum_{m \in \mathbb{Z}} u_m \delta^m.$$

Cette dernière somme étant une somme finie de suites car u est à support fini.

La linéarité du SLI T nous permet donc d'écrire

$$\begin{array}{rcl} T(u) & = & \sum_{m \in \mathbb{Z}} u_m T(\delta^m) \\ & = & \sum_{m \in \mathbb{Z}} u_m h^m \end{array}$$

Si on note $v = T(u)$, la dernière équation s'interprète comme une somme des suites h^m (m-translatée de h) pondérées chacune par un facteur u_m. Le terme v_n s'écrit donc

$$v_n = \sum_{m \in \mathbb{Z}} u_m h_{n-m}$$

car le terme numéro n dans chaque suite $u_m h^m$ (c.-à-d. la constante u_m multipliée par la suite h^m) est $u_m h_n^m = u_m h_{n-m}$ (par définition de la m-translatée).

Il nous reste à montrer que

$$\sum_{m \in \mathbb{Z}} u_m h_{n-m} = \sum_{l \in \mathbb{Z}} h_l u_{n-l}$$

Pour cela, il suffit d'appliquer le changement de variable

$$l = n - m.$$

□

Nous allons formaliser l'opération qui combine h et u pour obtenir $v = T(u)$. Elle s'appelle convolution. Vous pouvez voir sur la figure 1.10, une illustration graphique de la démonstration que nous venons de faire.

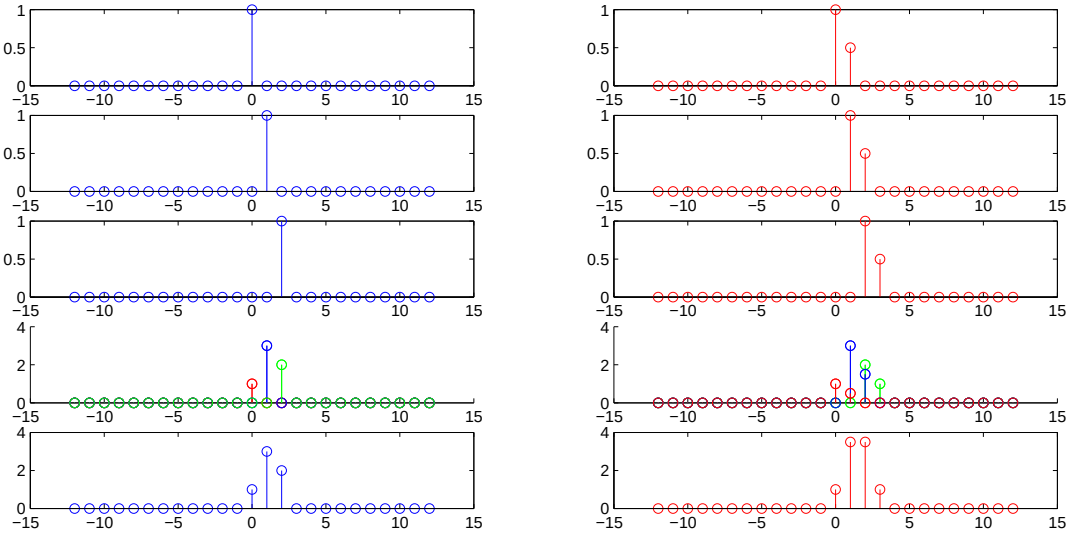


FIGURE 1.10 – **Ici on illustre à la fois le mécanisme de la démonstration de la proposition 1.8 et l'opération de convolution. Première ligne : gauche : l'impulsion en zéro (suite δ). Droite : la suite h. Deuxième ligne : la suite δ^1 (impulsion décalée de 1 vers la droite) puis la sortie qui lui est associée par T. L'invariance par translation impose que cette sortie soit h^1. Troisième ligne : même chose que ci-dessus, mais avec un décalage de 2. Quatrième ligne : On décompose une suite en trois pics, rouge vert et bleu, puis on trace sur un même graphique (à droite) la sortie associée à chaque pic. Dernière ligne : Par linéarité, la sortie associée à la suite de gauche est la somme des suites associées à chaque pic.**

Définition 1.9. La convolution

Si u et v sont deux suites, on appelle produit de convolution, et note $u * v$, la suite définie par

$$(u * v)_n = \sum_{m\in\mathbb{Z}} u_m v_{n-m} = \sum_{l\in\mathbb{Z}} v_l u_{n-l}$$

Sous réserve que la somme converge. Les règles de convergence de la somme infinie sont données au chapitre suivant. La seconde égalité est obtenue en faisant le changement de variable $l = n - m$. Elle implique que la **la convolution est commutative** *:*

$$u * v = v * u$$

Proposition 1.10. Propriétés de la convolution

on fixe trois suites u, v et w telles que les opérations de convolutions aient toujours un sens (les sommes infinies de la définition précédente convergent) alors on a

1. **Commutativité :** $u * v = v * u$
2. **Associativité :** $(u * v) * w = u * (v * w)$
3. **Linéarité :** $u * (v + w) = (u * v) + (u * w)$ *et* $\forall \lambda \in \mathbb{C}, u * (\lambda v) = \lambda(u * v)$
4. **Invariance par décalage :** *Si m est un entier et que w est la m-translatée de v, alors $u * w$ est la m-translatée de $u * v$. Autrement dit translater l'un des termes du produit de convolution translate le résultat d'une même distance (m, ici).*

Démonstration.

1. **Commutativité :** déjà vu, par le changement de variable $l = n - m$.
2. **Associativité :** Nous faisons les démonstrations dans le cas des suites à support fini. Si donc, u, v et w sont à support fini, les sommes que nous allons écrire sont finies et aucune justification n'est à donner quant à des interversions d'ordre de sommation

$$\forall n \in \mathbb{Z},\ ((u*v)*w)_n = \sum_{m\in\mathbb{Z}} (u*v)_m w_{n-m} = \sum_{m\in\mathbb{Z}} \left(\sum_{l\in\mathbb{Z}} u_l v_{m-l} \right) w_{n-m}$$

$$= \sum_{(l,m)\in\mathbb{Z}^2} u_l v_{m-l} w_{n-m} = \sum_{l\in\mathbb{Z}} \left(\sum_{m\in\mathbb{Z}} w_{n-m} v_{m-l} \right) u_l = \sum_{l\in\mathbb{Z}} \left(\sum_{r\in\mathbb{Z}} w_{n-l-r} v_r \right) u_l$$

La dernière égalité est obtenue par le changement de variable $r = m-l$. On applique maintenant la définition de la convolution pour obtenir

$$= \sum_{l\in\mathbb{Z}} (v*w)_{n-l} u_l = u*(v*w) \text{ on rappelle que } w*v = v*w$$

3. **Linéarité :**

$$(u*(v+w))_n = \sum_{m\in\mathbb{Z}} u_m \left(v+w\right)_{n-m} = \sum_{m\in\mathbb{Z}} u_m \left(v_{n-m} + w_{n-m}\right)$$

$$= \sum_{m\in\mathbb{Z}} (u_m v_{n-m}) + (u_m w_{n-m}) = \sum_{m\in\mathbb{Z}} u_m v_{n-m} + \sum_{m\in\mathbb{Z}} u_m w_{n-m} = (u*v)_n + (u*w)_n$$

Et, si $\lambda \in \mathbb{C}$

$$(u*(\lambda v))_n = \sum_{m\in\mathbb{Z}} u_m \left(\lambda v_{n-m}\right) = \lambda \sum_{m\in\mathbb{Z}} u_m v_{n-m} = \lambda (u*v)_n$$

4. **Invariance par translation :** On fixe un entier l et on note v^l la suite v translatée de l ($v^l_n = v_{n-l}$). On note $(u*v)^l$ la suite $(u*v)$ translatée de l. Ce qu'il faut montrer est

$$(u*v^l) = (u*v)^l$$

Développons le terme numéro n de $(u*v^l)$

$$(u*v^l)_n = \sum_{m\in\mathbb{Z}} u_m v^l_{n-m} = \sum_{m\in\mathbb{Z}} u_m v_{n-m-l} = \sum_{m\in\mathbb{Z}} u_m v_{(n-l)-m} = (u*v)_{n-l} = (u*v)^l_n$$

□

Proposition 1.11. SLI définis par une convolution
Soit V et W deux espaces invariants par translation et h une suite telle que

$$\forall u \in V, h*u \in W$$

Alors la fonction T définie de V vers W par

$$\forall u \in V, T(u) = h*u$$

est un SLI et h est la **réponse impulsionnelle** *du SLI T.*

Démonstration. Il suffit d'appliquer les propriétés de linéarité et d'invariance par décalage de la convolution vue à la proposition précédente. □

Définition 1.12. Définitions de certains espaces invariants par translation

1. *Une suite est dite* **bornée** *si*

$$\exists A \in \mathbb{R}_+,\ \forall n \in \mathbb{Z},\ |u_n| \leqslant A$$

2. *Une suite est dite* **sommable** *si*

$$\sum_{n\in\mathbb{Z}} |u_n| < +\infty$$

3. *Une suite est dite d'***énergie finie**

$$\sum_{n\in\mathbb{Z}} |u_n|^2 < +\infty$$

Les espaces des suites bornées, sommables ou d'énergie finie sont des espaces stables par translation.

Théorème 1.13. Tous les SLI sont des convolutions
Si T est un SLI entre des espaces V et W alors il existe une suite h telle que

$$\forall u \in V, T(u) = u * h$$

Pour ce cours on admettra cette affirmation. Elle est vraie dans de nombreux cas d'espaces V et W. Citons quelques cas

1. *$V = W$ =ensemble des suites bornées et T vérifie la condition technique*[4]

$$\forall v \in V, \lim_{N\to+\infty} T(v^N) = 0 \text{ où la suite } v^N \text{ est définie par } v_n^N = \begin{cases} 0 \text{ si } |n| < N \\ v_n \text{ sinon} \end{cases}$$

   alors il existe une suite h sommable ($\sum_{n\in\mathbb{Z}} |h_n| < +\infty$)
   telle que

$$\forall u \in V, T(u) = h * u$$

2. *Si V est l'espace des suites d'énergie finie et W l'espace des suites bornée, alors*

$$\exists h \in V, \forall u \in V, T(u) = h * u$$

   Ainsi les SLI qui agissent sur l'espace des suites d'énergie finie ont une réponse impulsionnelle qui est elle-même d'énergie finie.

3. *Si $V = W =$ l'espace des suites sommables alors*

$$\exists h \in V, \forall u \in VT(u) = h * u$$

   Là encore les SLI définis sur les suites sommables et à sortie sommable ont une réponse impulsionnelle sommable.

4. Cette condition est là pour éviter des cas pathologiques qui n'ont aucune réalité physique. Elle signifie que T n'a pas de "masse à l'infini". Elle est là par souci de rigueur.

4. Si V est l'espace des suites sommables et W celui des suites bornées alors

$$\exists h \in W \ \forall u \in V, T(u) = h * u$$

Les SLI qui transforment une suite sommable en une suite bornée ont une réponse impulsionnelle bornée

Toutes les démonstrations de ces cas particuliers se font de la même manière que pour les suites à support fini en ajoutant des considérations de convergence dans les espaces de dimension infinie que sont V et W. Elles ne sont pas faites ici.

Exemple 1.14. **Variations d'une série temporelle**
Si u est une suite bornée et que l'on définit la suite $v = T(u)$ par

$$v_n = u_n - u_{n-1}$$

alors on voit facilement que v est aussi une suite bornée. On vérifie aussi que

$$T(u) = h * u$$

où la réponse impulsionnelle h est définie par

$$h_0 = 1, \ h_1 = -1, \text{ et } h_n = 0 \text{ dans les autres cas}$$

Exemple 1.15. **Moyenne glissante**
Reprenons l'exemple de la moyenne glissante. On avait, si u est la valeur journalière d'une grandeur (température, cours de bourse...) et v sa moyenne glissante sur le mois précédent,

$$v_n = \frac{1}{30}\left(u_n + u_{n-1} + u_{n-2} + \cdots + u_{n-29}\right)$$

On voit que v est le résultat de convolution de u avec la suite h définie par :

$$h_n = \begin{cases} \frac{1}{30} \text{ si } 0 \leqslant n \leqslant 29 \\ 0 \text{ sinon} \end{cases}$$

1.2.3 Les ondes harmoniques ou ondes de Fourier

Dans cette partie nous cherchons des suites particulières qui seraient modifiées très simplement par un SLI. On cherche des suites, qui après passage par un SLI sortiraient inchangées, ou tout au plus multipliées par une constante.

On veut une suite u telle que
— Pour tout SLI T $\exists C \in \mathbb{C}, \ T(u) = Cu$.
— La suite u doit rester bornée (afin qu'elle ait un sens physique).
— On fixe $u_0 = 1$ (car les deux hypothèses ci-dessus sont inchangées par une multiplication par une constante, on élimine ainsi un degré de liberté trivial).

Une telle suite est appelée suite harmonique ou onde de Fourier. On a le résultat suivant :

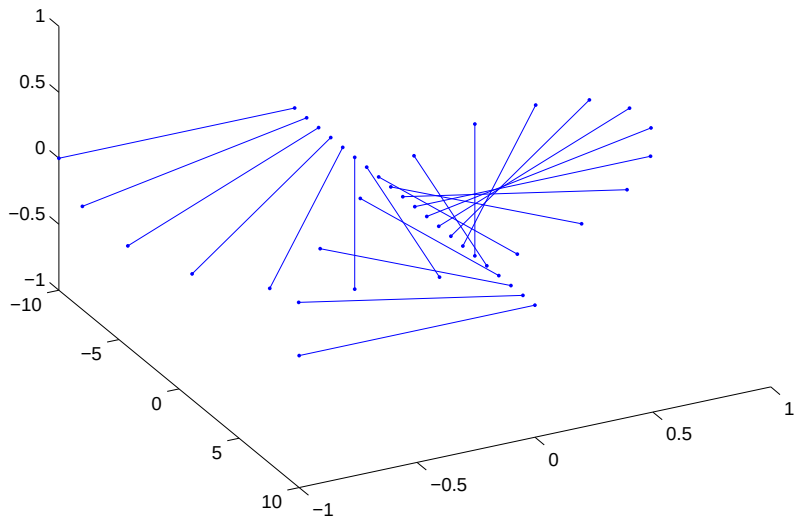


FIGURE 1.11 – **Une onde de Fourier sur $\mathbb{Z}$. Nous l'avons représentée en trois dimensions. L'axe $y = z = 0$ est celui de l'index de la suite. Il est matérialisé par les points alignés. Nous avons ensuite joint chaque point de cet axe au point $x = n, y = cos(2i\pi\nu n), z = sin(2i\pi\nu n)$.**

Proposition 1.16. Caractérisation des ondes de Fourier sur $\mathbb{Z}$
Une suite vérifie les conditions ci-dessus si et seulement si :

$$\exists \nu \in [-1/2, 1/2[\text{ tel que } u_n = e^{2i\pi\nu n}$$

On appelle ν la fréquence de la suite harmonique u. Si ν convient dans la formule ci-dessus alors tout $\nu + m$ où $m \in \mathbb{Z}$ convient aussi. (La figure 1.11 montre une onde de Fourier sur $\mathbb{Z}$)

Démonstration. Considérons le SLI suivant :

$$T(v)_n = v_{n-1}$$

Il s'agit du SLI qui décale une suite d'une distance de 1.

D'après les hypothèses faites sur u on doit avoir

$$\exists C \in \mathbb{C},\ u_{n-1} = Cu_n$$

la constante C ne peut être nulle car on a

$$1 = u_0 = Cu_1$$

Comme, de plus, on a $u_0 = 1$, on obtient, par récurrence,

$$\forall n \in \mathbb{Z}, u_n = C^n$$

Comme u est supposée bornée, on doit avoir

$$|C| = 1.$$

car sinon $|u_n|$ tendrait vers $+\infty$ lorsque n tend vers $\pm\infty$ (suivant que $|C| > 1$ ou $|C| < 1$).

Donc il existe $\nu \in [-1/2, 1/2[$ telle que

$$C = e^{2i\pi\nu}$$

Et enfin

$$\forall n \in \mathbb{Z},\ u_n = e^{2i\pi\nu n}$$

Nous avons montré que, si u vérifie les hypothèses d'onde harmonique, alors il est de la forme ci-dessus. Il nous faut maintenant démontrer que si u est de la forme ci-dessus alors il vérifie les conditions de suite harmonique.

Soit donc T un SLI (qui peut prendre u en entrée) et u de la forme

$$u_n = e^{2i\pi\nu n}$$

On a déjà u bornée et $u_0 = 1$. Il reste à montrer que

$$\exists C \in \mathbb{C},\ T(u) = Cu$$

Notons u^m la m-translatée de u. On a, par la formule donnant u

$$u_n^m = e^{-2i\pi\nu m} e^{2i\pi\nu n} = e^{-2i\pi\nu m} u_n$$

Ce qui s'écrit

$$u^m = e^{-2i\pi\nu m} u$$

(translater u revient à la multiplier par la constante $e^{-2i\pi\nu m}$)

Pour alléger les notations, on note $v = T(u)$ et v^m la m-tanslatée de v. Les propriétés des SLI font que

$$v_{-m} = (v^m)_0 = T(u^m)_0 = T(e^{-2i\pi\nu m} u)_0 = e^{-2i\pi\nu m}(T(u))_0 = v_0 e^{2i\pi\nu(-m)} = v_0 u_{-m}$$

La seconde égalité est l'expression de l'invariance par translation d'un SLI. La quatrième est l'expression de la linéarité des SLI. Les autres ne sont que des réécritures des termes qui se trouvent à leur gauche.

Ceci étant vrai pour tout m on a

$$v = v_0 u$$

la constante C que l'on cherche est v_0.

Dans cette démonstration, le point qui a joué le plus grand rôle est le fait qu'une onde harmonique (ou de Fourier) translatée est égale à elle même multipliée par une constante. □

Les ondes harmoniques sont déterminées par leur fréquence (ν dans les formules ci-dessus). Pour un SLI fixé, chaque onde de fréquence $\nu \in [-1/2, 1/2[$ se trouve multipliée (lorsqu'elle passe dans le SLI) par une constante C qui dépend de ν. Nous noterons donc $C(\nu)$ la valeur de la constante pour chaque fréquence.

Proposition 1.17. Gain fréquentiel, définition et valeur
Soit T un SLI (qui admet les ondes de Fourier en entrée) et h sa réponse impulsionnelle. Pour chaque suite harmonique de fréquence ν notée u^ν ($u^\nu_n = e^{2i\pi\nu n}$) on sait par ce qui précède que

$$\exists C(\nu) \in \mathbb{C}, T(u^\nu) = C(\nu)u^\nu$$

$C(\nu)$ est appelé **gain fréquentiel** *de T et sa valeur est donnée par*

$$C(\nu) = \sum_{n\in\mathbb{Z}} h_n u^\nu_{-n} = \sum_{n\in\mathbb{Z}} h_n e^{-2i\pi\nu n}.$$

Démonstration. Si on note $v^\nu = T(u^\nu)$ (la sortie associée à l'onde harmonique) on a d'une part

$$v^\nu_0 = C(\nu)u^\nu_0 = C(\nu)$$

(par la définition du gain fréquentiel)

et d'autre part, la caractérisation des SLI par leur réponse impulsionnelle nous donne,

$$v^\nu_0 = \sum_{n\in\mathbb{Z}} h_n u^\nu_{-n} = \sum_{n\in\mathbb{Z}} h_n e^{-2i\pi\nu n}$$

□

Remarque 1.18. Le fait que les $C(\nu)$ suffisent à caractériser le SLI sera vu au chapitre transformée de Fourier.

Ici on veut calculer $C(\nu)$ pour les deux SLI, moyenne glissante et variations journalières.

Exemple 1.19. **Moyenne glissante**
Pour toute fréquence ν, on a (le calcul ci-dessous est valable pour $\nu \neq 0$, c.-à-d. $e^{2i\pi\nu} \neq 1$)

$$C(\nu) = v_0 = \frac{1}{30}(e^{2i\pi 0\nu} + e^{2i\pi.-1\nu} + \cdots + e^{2i\pi.-29\nu}) = \frac{1}{30}\frac{1-e^{-2i\pi\nu 30}}{1-e^{-2i\pi\nu}} = \frac{1}{30}e^{-i\pi.29\nu}\frac{sin(30\pi\nu)}{sin(\pi\nu)}$$

Pour $\nu = 0$

$$C(0) = v_0 = \frac{1}{30}(1+1+\cdots+1) = 1$$

On trace le graphique de $|C(\nu)|$ (figure 1.12) et on constate bien que les ondes de basse fréquence (autour de $\nu = 0$) sont privilégiées par rapport aux hautes fréquences (autour de $-1/2$ et $1/2$). Ce SLI est un passe-bas.

Exemple 1.20. **Variations journalières**
Dans le cas des variations journalières, on a la formule :

$$C(\nu) = v_0 = 1 - e^{-2i\pi\nu} = 2ie^{-i\pi\nu}sin(\pi\nu)$$

Le graphique 1.13 montre le module de $C(\nu)$ en fonction de ν. On constate bien que les hautes fréquences (autour de $-1/2$ et $1/2$) sont privilégiées par rapport aux basses fréquences (autour de la fréquence 0). Ce SLI est un passe-haut.

Remarque 1.21. Les constantes $C(\nu)$ sont caractéristiques de chaque SLI. On les nomme gain fréquentiel. Elles suffisent à décrire entièrement un SLI (ce que l'on n'a pas prouvé ici). On verra au chapitre suivant qu'elles sont la transformée de Fourier de la réponse impulsionnelle. Nous disposons ainsi de deux caractérisations duales des SLI

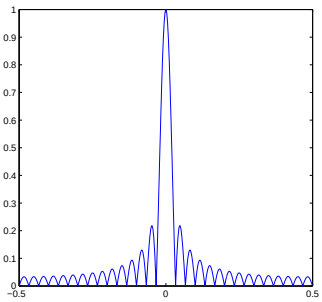


FIGURE 1.12 – **Module de $|C(\nu)|$ pour un filtre moyenne glissante. Remarquer comme les ondes de basse fréquence (autour de 0) sont privilégiées.**

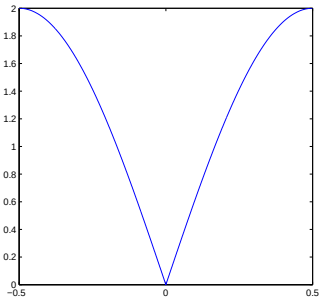


FIGURE 1.13 – **Module de $|C(\nu)|$ pour le filtre des variations journalières. Remarquer comme les ondes de haute fréquence (autour de $|\nu| = 1/2$) sont privilégiées.**

1. La réponse impulsionnelle qui définit le SLI par l'intermédiaire de l'opération de convolution.
2. La réponse en fréquence qui définit le SLI par l'intermédiaire de son action sur les ondes de Fourier et qui fait l'objet du chapitre suivant.

Dans la suite nous allons étendre les notions vues pour les signaux définis sur $\mathbb{Z}$ à d'autres types de signaux, ceux définis sur $\mathbb{R}$, sur $\{0, \ldots, N-1\}$ et les signaux périodiques (que l'on représente par des fonctions définies sur $[-1/2, 1/2[$).

Nous nous contentons de donner les définitions. La démonstration du fait que tout SLI est une convolution par une réponse impulsionnelle est soit identique à celle faite sur $\mathbb{Z}$, soit fait intervenir des outils mathématiques que nous n'avons pas à notre disposition.

Vous remarquerez que les sections qui suivent sont très répétitives. C'est un choix destiné à montrer que les SLI ont une structure algébrique universelle (la convolution) et un comportement qui peut se décrire suivant leur action sur les ondes pures (gain fréquentiel). La répétition des définitions et des propriétés vous permettra d'intégrer ce qu'il y a d'intrinsèque dans les SLI et de ne retenir qu'un petit nombre de propriétés génériques qui vous permettront de manipuler les SLI dans tous les cas abordés dans ce cours.

1.3 Les SLI pour les signaux en temps continu

Définition 1.22. Translation d'une fonction définie sur $\mathbb{R}$
Pour toute fonction f définie sur $\mathbb{R}$ à valeurs complexes, et tout réel x on définit la x-translatée de f, que l'on peut noter f_x par

$$\forall y \in \mathbb{R}, f_x(y) = f(y - x)$$

Définition 1.23. Espaces de fonctions stables par translation
Un sous espace vectoriel, V, de l'espace des fonctions définies sur $\mathbb{R}$ est dit stable par translation si

$$\forall f \in V, \forall x \in \mathbb{R}, f_x \in V$$

Des exemples d'espaces invariants par translation seront vus au chapitre suivant.

Définition 1.24. SLI pour les fonctions définies sur $\mathbb{R}$
Si V et W sont deux sous-espaces de fonctions, stables par translation et T une application de V vers W. On dit que T est un SLI si

1. *T est linéaire ($T(f+g) = T(f) + T(g)$ et $T(\lambda f) = \lambda T(f)$)*
2. *T est invariant par translation*

$$\forall f \in V, x \in \mathbb{R},\ T(f_x) = T(f)_x$$

où $T(f)_x$ est la x-translatée de la fonction $T(f)$.

Définition 1.25. Convolution de fonctions
Si f et g sont deux fonctions définies sur $\mathbb{R}$, on appelle produit de convolution de f et g que l'on note $f * g$, la fonction qui vaut (lorsque l'intégrale a un sens)

$$\forall x \in \mathbb{R}, (f * g)(x) = \int_{t \in \mathbb{R}} f(t)g(x - t)dt = \int_{z \in \mathbb{R}} g(z)f(x - z)dz$$

Proposition 1.26. Propriétés de la convolution
on fixe trois fonctions f, g et h telles que les opérations de convolutions aient toujours un sens (les intégrales de la définition précédente convergent) alors on a :

1. **Commutativité :** $f * g = g * f$
2. **Associativité :** $(f * g) * h = f * (g * h)$
3. **Linéarité :** $f * (g + h) = (f * g) + (f * h)$ *et* $\forall \lambda \in \mathbb{C}, f * (\lambda g) = \lambda(f * g)$
4. **Invariance par décalage :** *Si x est un réel et que g_x est la x-translatée de g, alors $f * g_x$ est la x-translatée de $f * g$. Autrement dit translater l'un des termes du produit de convolution translate le résultat d'une même distance (x, ici).*

Démonstration.
Les preuves sont soient aussi simples que dans le cas des suites à support fini, soit font intervenir des théorèmes d'intégration non vus. □

Définition 1.27. SLI définis par une convolution
Soient V et W deux espaces de fonctions stables par translation. Soit h une fonction telle que

$$\forall f \in V, h * f \in W$$

On définit l'application T de V vers W par

$$\forall f \in V, T(f) = h * f.$$

Alors, T est un SLI et h est appelé **réponse impulsionnelle** *de T.*

Démonstration.
Ayant pris soin de s'assurer que l'opération de convolution par h transformait bien une fonction de V en une fonction de W, la démonstration est purement formelle et se recopie de celle faite pour les SLI sur $\mathbb{Z}$. □

Théorème 1.28. Tous les SLI sont des convolutions
Si T est un SLI entre deux espaces V et W alors il existe une fonction h telle que

$$\forall f \in V, T(f) = f * h$$

Démonstration. **Justification de l'absence de démonstration**
Un ingrédient de la démonstration faite à la proposition 1.8 était l'utilisation de la suite impulsion en zéro

$$\delta_n^0 = \begin{cases} 1 \text{ si } n = 0 \\ 0 \text{ si } n \neq 0 \end{cases}$$

Or l'équivalent sur $\mathbb{R}$ d'une telle impulsion est la fameuse distribution de Dirac. Comme nous ne possédons pas cet instrument mathématique (les distributions en général), nous ne faisons pas la démonstration. □

Définition 1.29. Ondes de Fourier sur $\mathbb{R}$
On appelle onde de Fourier sur $\mathbb{R}$ toute fonction f telle que

$$\exists \nu \in \mathbb{R}\ \forall x \in \mathbb{R}, f(x) = e^{2i\pi\nu x}$$

Le paramètre ν est appelé **fréquence** *de l'onde f. (La figure 1.14 représente une telle onde de Fourier)*

Remarque 1.30.
Par rapport au cas de $\mathbb{Z}$, la fréquence ν peut être n'importe quel réel et pas seulement dans l'intervalle $[-1/2, 1/2[$. De plus, l'onde de fréquence $1 + \nu$ n'est pas égale à l'onde de fréquence ν, comme c'était le cas pour $\mathbb{Z}$. Cette différence fondamentale jouera un rôle crucial dans le théorème d'échantillonnage (voir plus loin).

Sur $\mathbb{Z}$ nous avions caractérisé les ondes de Fourier par le fait qu'elles sont des vecteurs propres de tout SLI et en avions déduit leur forme analytique. Nous aurions pu faire la même chose ici en ajoutant la contrainte qu'elles doivent être continues. Refaites le raisonnement en exercice dans le cas de $\mathbb{R}$ (on pourra remarquer que l'opération translation de 1 est un SLI sur l'ensemble des fonctions bornées).

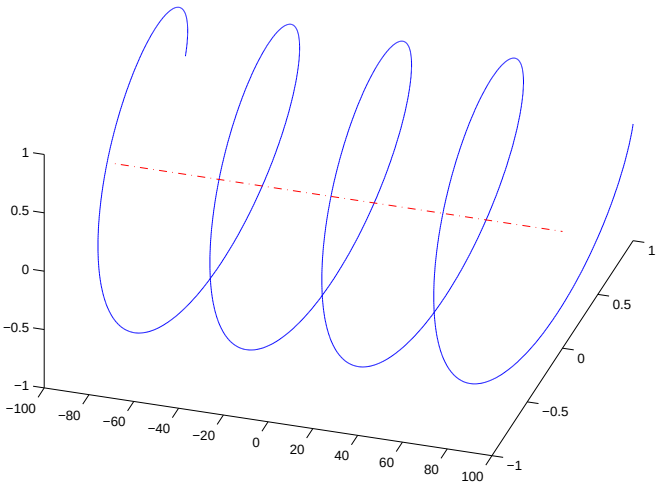


FIGURE 1.14 – **Une onde de Fourier sur $\mathbb{R}$. Nous l'avons représentée en trois dimensions. L'axe $y = z = 0$ est celui du paramètre de l'onde. En prenant comme paramètre $t \in \mathbb{R}$, la courbe représente les points de la forme $x = t, y = cos(2i\pi t\nu), z = sin(2i\pi t\nu)$.**

Proposition 1.31. *Si T est un SLI sur $\mathbb{R}$ (qui accepte en entrée les ondes de Fourier) et f une onde de Fourier alors*

$$\exists C \in \mathbb{C}, T(f) = Cf.$$

Autrement dit, une onde de Fourier est simplement multipliée par une constante (complexe) lorsqu'elle passe dans un SLI. Cette constante peut dépendre de la fréquence.

Démonstration.
Toute onde de Fourier, f, vérifie (voir sa formule)

$$f(x+y) = f(x)f(y) \text{ et } f(-x) = \frac{1}{f(x)} = \overline{f(x)}$$

Et on a (on rappelle que f_x est la x-translatée de f)

$$[T(f)](x) = [T(f)_{-x}](0) = [T(f_{-x})](0) = [T\left(t \mapsto f(x)f(t)\right)](0) = f(x)[T(t \mapsto f(t))](0) = Cf(x)$$

avec $C = (T(f))(0)$. □

Définition 1.32. Réponse fréquentielle d'un SLI
Si T est un SLI qui admet les ondes de Fourier en entrée, alors on appelle réponse en fréquence (ou gain fréquentiel) de T la fonction C définie sur $\mathbb{R}$ telle que

$$\forall \nu \in \mathbb{R}, T(f^\nu) = C(\nu)f^\nu$$

où f^ν est l'onde de Fourier de fréquence ν (i.e. $f^\nu(x) = e^{2i\pi\nu}$).

On verra dans le chapitre suivant que C est la transformée de Fourier de la réponse impulsionnelle. Comme pour le cas discret C suffit a caractériser un SLI et on a encore deux descripteurs duaux pour chaque SLI : Sa réponse impulsionnelle et sa réponse en fréquence.

Exemple 1.33. **Évolution de la distribution de température**
Sur une barre métallique assez longue pour être considérée comme infinie, on suppose qu'à l'instant $t = 0$, la température a une distribution donnée par la fonction $f(x, 0)$ (de $x \in \mathbb{R}$ vers $\mathbb{R}$). On sait que la température évolue suivant l'équation :

$$\frac{\partial f}{\partial t} = \lambda \frac{\partial^2 f}{\partial x^2}$$

Soit un temps $t_0 > 0$, comment déduire la distribution de température au temps t_0, (*i.e* $x \mapsto f(x, t_0)$ à partir de la distribution initiale($x \mapsto f(x, 0)$) ? Nous allons voir que cette relation est un SLI.

Si on note $h(x) = f(x, 0)$ et $g(x) = f(x, t_0)$ et qu'on les considère comme l'entrée (h) et la sortie (g) d'un système. Il est clair que ce système est invariant par rapport à une translation de la variable x. En effet, l'équation qui gouverne la diffusion de la chaleur est invariante par translation par rapport à la variable x. De plus, la relation qui lie g à h est linéaire, car si $f_1(x, t)$ et $f_2(x, t)$ sont solutions de l'équation de la chaleur alors $f_1 + f_2$ est aussi solution de l'équation (il faut ajouter à cela des théorèmes d'unicité de la solution de l'équation différentielle).

Nous avons dit que les SLI sont presque toujours des convolutions. On s'attend donc à ce qu'il existe une fonction r_{t_0} telle que

$$g = r_{t_0} * h.$$

Et c'est bien le cas. Il suffit de prendre

$$r_{t_0}(x) = \frac{1}{2\sqrt{2\pi}\lambda t_0} e^{-\frac{x^2}{2\lambda t_0}}$$

Pour voir cela, il suffit de vérifier que la fonction à deux variables

$$f(x, t) = (r_t * h)(x), \text{ La convolution portant sur la variable } x$$

est solution de l'équation de la chaleur et conclure par unicité de la solution [5].

Exemple 1.34. **Enregistrement d'un son**
Dans une pièce se trouve un générateur de son qui émet un signal sonore noté f. On dispose également un enregistreur. Le signal que l'enregistreur enregistre est noté g. Le son émis se propage dans la pièce avant d'atteindre l'enregistreur. L'onde sonore peut se propager en ligne droite ou encore rebondir sur un mur. Au final les équations de l'acoustique montrent que g dépend de f de manière linéaire. De plus, la pièce étant invariante dans le temps, on imagine facilement que décaler f dans le temps entraîne l'enregistrement d'un son identique à celui enregistré sans décalage.

5. L'élève intéressé pourra démontrer que la fonction ci-dessus est bien solution en développant la fonction f(x,t) puis en dérivant sous le signe somme par rapport à t. Plus la règle que $(f*g)' = f'*g = f*g'$ où la dérivation porte sur la même variable que celle sur laquelle porte la convolution.

Le lien entre ce qu'enregistre un appareil et ce qui est émis est un SLI. Suivant les caractéristiques de la pièce (masse et dimensions des murs, position de la fenêtre et des meubles) il peut y avoir plus ou moins d'écho.

Une pièce faite pour l'enregistrement professionnel a ses parois recouvertes de matériaux absorbants afin de réduire les réverbérations et faire que g (le son enregistré) soit presque identique au son émis f.

Pour une pièce fixée, il existe une fonction h telle que

$$g = h * f$$

(car tout SLI est une convolution)

Certains systèmes hi-fi sont livrés avec des micros servant au calibrage de la pièce dans laquelle ils sont utilisés. L'utilisateur est invité à positionner le micro là où il souhaite écouter de la musique. Ensuite la calibration commence.

La chaîne Hi-Fi émet des sons de forte intensité (afin qu'ils couvrent les perturbations dues aux bruits extérieurs). L'ordinateur de la chaîne Hi-Fi connaît exactement le son émis. Elle compare le son enregistré par le micro et celui qu'elle a émis et en déduit la réponse impulsionnelle h. Connaître h c'est connaître la caractéristique pertinente de la pièce du point de vue de la transmission fidèle (ou pas) du son de la chaîne à l'utilisateur.

Exemple 1.35. **Inverser la réverbération ?**
Soit T un SLI qui donne, à partir d'un son émis en un point d'une pièce le son enregistré en un autre point. On peut s'intéresser à inverser T. Or l'inverse d'un SLI ne peut être qu'un SLI, car l'inverse d'une application linéaire est linéaire et de plus, si on note T^{-1} l'application inverse de T, et $g = T(f)$ alors

$$T^{-1}(g_x) = T^{-1}(T(f)_x)T^{-1}(T(f_x)) = f_x = [T^{-1}(g)]_x$$

Ainsi, on sait qui si une chaîne Hi-Fi veut inverser l'effet de réverbération d'une pièce, elle doit appliquer au signal sonore, avant de l'émettre, un filtrage SLI.

Comment calculer le SLI inverse d'un SLI, est une question à laquelle on répondra de diverses manières dans le reste du cours.

Remarque 1.36. Dans la suite, on continue à étendre nos définitions et propriétés classiques des SLI à deux nouveaux ensembles. Nous serons très brefs, et insisterons surtout l'opération de décalage qui a une forme circulaire sur les deux ensembles suivants, ce qui pourrait dérouter.

1.4 Les SLI pour les signaux finis périodiques

Définition 1.37. signaux finis périodiques *Les signaux finis périodiques sont les signaux définis sur un ensemble de la forme $\{0, \ldots, N-1\}$, où N est un entier strictement positif. On peut aussi noter $\mathbb{Z}/N\mathbb{Z}$ pour l'ensemble de définition de ces signaux. On les appelle périodiques, car toutes les définitions qui suivent feront l'hypothèse que l'échantillon d'index $N-1$ est proche de l'échantillon 0.*

Remarque 1.38. Le mot de périodique provient ici du fait qu'il est équivalent de se donner une suite périodique de période N et se donner un signal fini de taille N (que l'on répète sur tout $\mathbb{Z}$ pour obtenir le signal infini périodique d'origine).

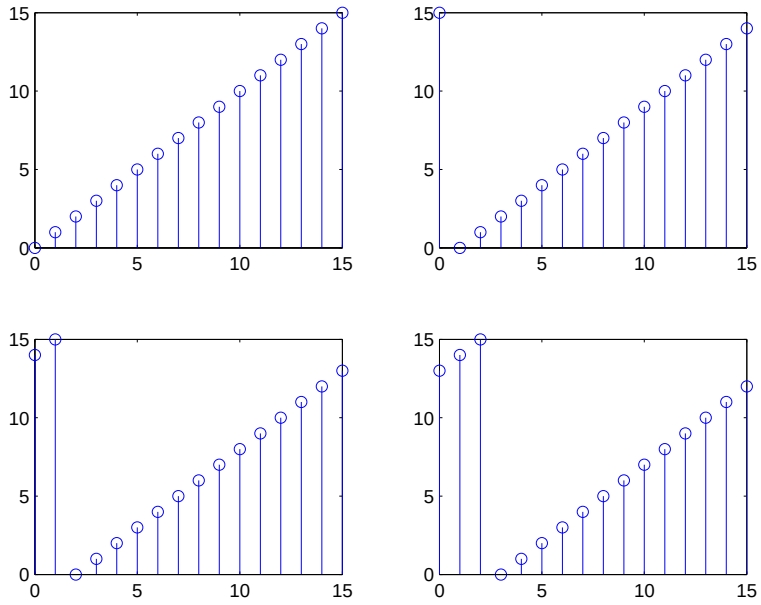


FIGURE 1.15 – **Haut à gauche : Le signal fini périodique. Haut droit : Ce même signal décalé d'un échantillon vers la droite. Remarquer comme le dernier échantillon est rentré par la gauche. Bas à droite : même signal décalé de deux échantillons vers la droite. Bas à droite : Décalage de trois échantillons.**

Définition 1.39. Opération d'addition dans $\{0, \ldots, N-1\}$
Si x et y sont deux éléments de $\{0, \ldots, N-1\}$ leur somme, dans cet ensemble, est définie comme étant la somme modulo N. Par exemple $(N-1)+2 = N+1 = 1$ et $1+2 = 3$ si $N > 3$.

Zéro est l'élément neutre pour l'addition et tout élément n de $\{0, \ldots, N-1\}$ possède un inverse qui est l'égal modulo N de l'entier relatif $-n$ ($\forall x \in \{0, \ldots, N-1\}$ $\exists y \in \{0, \ldots, N-1\}$ $x + y = 0$).

Définition 1.40. La translation "circulaire" d'un signal fini périodique
Si u est signal fini périodique et $m \in \{0, \ldots, N-1\}$ on appelle m-translatée de u la suite, que l'on note v, définie par

$$v_n = u_{n-m}$$

Le $n-m$ signifiant l'égal modulo N de l'entier relatif $n-m$. Un exemple de translations de signal fini périodique est donné à la figure 1.15.

Définition 1.41. Les SLI et convolutions circulaires
Toute application, T, linéaire et invariante par translation entre l'espace des signaux finis périodiques et lui-même, sont de la forme

$$T(u)_n = \sum_{m=0}^{m=N-1} u_m h_{n-m}$$

où h est appelé réponse impulsionnelle de T. L'opération entre h et u pour obtenir $T(u)$ est appelée convolution. L'opération de convolution a les mêmes propriétés que sur les autres espaces.

Autrement dit, tous les SLI sur les signaux fini périodiques sont des convolutions et toute convolution contre une réponse impulsionnelle est un SLI.

Définition 1.42. Les ondes de Fourier correspondantes
Les ondes de Fourier ont la propriété que, si ϕ est une onde de Fourier,

$$\forall x, y \in \{0, \dots, N-1\}\ \phi(x+y) = \phi(x)\phi(y)$$

Avec cette seule condition, on a

$$\exists k \in \{0, \dots, N-1\},\ \forall n \in \{0, \dots, N-1\} \phi(n) = e^{2i\pi \frac{k}{N} n}$$

Le réel $\frac{k}{N}$ est appelé fréquence de l'onde de Fourier ϕ.

Comme dans les autres cas, si T est un SLI et que u est une onde de Fourier (fini discret) de fréquence ν alors il existe une complexe $C(\nu)$ appelé gain fréquentiel ou réponse fréquentielle tel que

$$T(u) = C(\nu)u$$

Remarque 1.43. Par rapport aux espaces déjà vus ($\mathbb{Z}$ et $\mathbb{R}$) les fréquences sont maintenant quantifiées (elles sont de la forme $\frac{k}{N}$). En effet, si ϕ est une onde de Fourier, on sait que $\phi(Nx) = \phi(x)^N$. Or, pour tout $x \in \{0, \dots, N-1\},\ Nx = 0$ (modulo N) on a donc

$$\phi(x)^N = 1 (= \phi(0))$$

La figure 1.16 montre une onde de Fourier sur $\{0, \dots, N-1\}$ pour $N = 16$. Sa fréquence est $2/16$.

1.5 Les SLI pour les signaux périodiques

Définition 1.44. signaux 1-périodiques *Les signaux 1-périodiques (ou périodiques, pour simplifier les formulations par la suite) sont les signaux définis sur l'ensemble $[-\frac{1}{2}, \frac{1}{2}[$.*

Remarque 1.45. Comme à la section précédente le mot de périodique provient ici du fait qu'il est équivalent de se donner une fonction 1-périodique et se donner un signal défini sur $[-\frac{1}{2}, \frac{1}{2}[$ (que l'on répète sur tout $\mathbb{R}$ pour obtenir le signal infini en temps continu).

Définition 1.46. Opération d'addition dans $[-\frac{1}{2}, \frac{1}{2}[$
Si x et y sont deux éléments de $[-\frac{1}{2}, \frac{1}{2}[$ leur somme, dans cet ensemble, est définie comme étant la somme modulo 1. Par exemple $0.3+0.3 = 0.6 = -0.4+1 = -0.4$, $0.1+0.2 = 0.3$, et $-0.3 + 0.5 = 0.2$.

Zéro est l'élément neutre pour l'addition et tout élément x de $[-\frac{1}{2}, \frac{1}{2}[$ possède un inverse $-x$ (sauf $-1/2$ qui est son propre inverse).

Définition 1.47. La translation "circulaire" d'un signal périodique
Si f est signal périodique et $x \in [-\frac{1}{2}, \frac{1}{2}[$ on appelle x-translatée de f la fonction, que l'on note f_x, définie par

$$\forall y \in [-\frac{1}{2}, \frac{1}{2}[,\ f_x(y) = f(y-x) = \begin{cases} f(y-x) \text{ si } y \geqslant x \\ f(y-x+1) \text{ si } y < x \end{cases}$$

(les expressions dans l'accolade utilisent les règles d'addition sur $\mathbb{R}$ d'où le $y - x + 1$ pour signifier que $y - x$ va sortir de l'intervalle $[-\frac{1}{2}, \frac{1}{2}[$)

Un exemple de translations de signal périodique est donné à la figure 1.17.

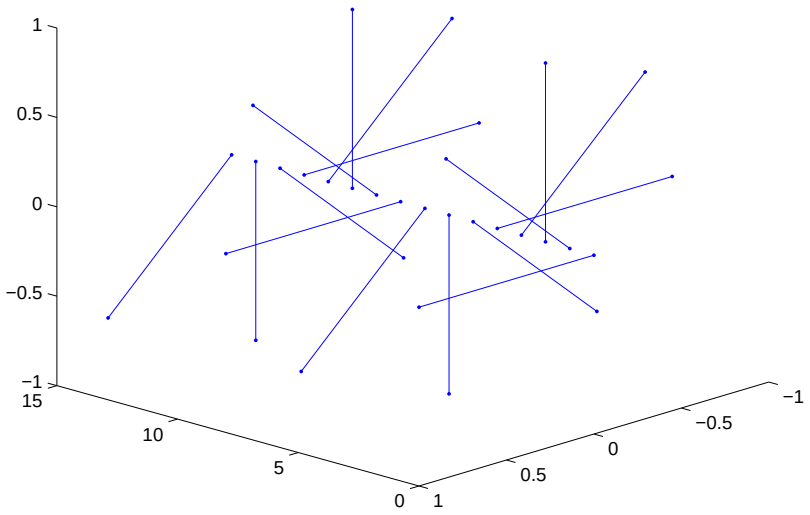


FIGURE 1.16 – **Une onde de Fourier sur $\{0 \dots 15\}$ de fréquence $2/16$. Noter comme elle parcourt exactement deux périodes le long de son domaine de définition.**

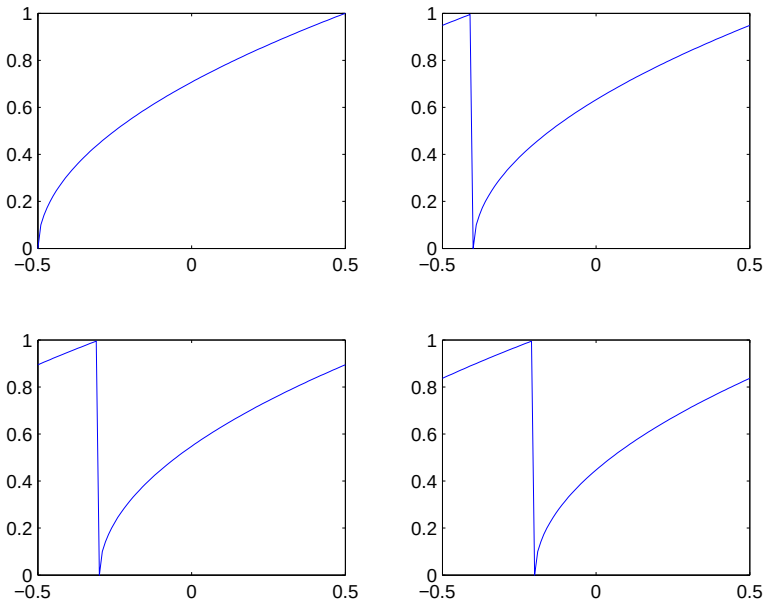


FIGURE 1.17 – **Haut à gauche : Le signal périodique. Haut droit : Ce même signal décalé de 0.1 vers la droite. Remarquer comme la dernière partie est entrée par la gauche du graphe. Remarquer aussi que cette fonction n'est pas continue si on la regarde comme une fonction périodique définie sur $\mathbb{R}$. Bas à droite : même signal décalé de 0.2 vers la droite. Bas à droite : Décalage de 0.3.**

Définition 1.48. Les SLI et convolutions circulaires
Toute application, T, linéaire et invariante par translation entre l'espace des signaux 1-périodiques et lui-même, est de la forme

$$[T(f)](x) = \int_{-\frac{1}{2}}^{\frac{1}{2}} f(t)h(x-t)dt = \begin{cases} \int_{-\frac{1}{2}}^{-\frac{1}{2}+x} f(t)h(x-t-1) + \int_{-\frac{1}{2}+x}^{\frac{1}{2}} f(t)h(x-t) \text{ si } x \geqslant 0 \\ \\ \int_{-\frac{1}{2}}^{\frac{1}{2}+x} f(t)h(x-t) + \int_{\frac{1}{2}+x}^{\frac{1}{2}} f(t)h(x-t+1) \text{ si } x \leqslant 0 \end{cases}$$

(les opérations d'addition et de soustraction dans les expressions après accolade sont des opérations dans $\mathbb{R}$ qui garantissent le retour dans l'intervalle $[-\frac{1}{2}, \frac{1}{2}[$)

où h est appelé réponse impulsionnelle de T. L'opération entre h et f pour obtenir $T(f)$ est appelée convolution. L'opération de convolution a les mêmes propriétés que sur les autres espaces.

Autrement dit, tous les SLI sur les signaux périodiques sont des convolutions et toute convolution contre une réponse impulsionnelle est un SLI.

Définition 1.49. Les ondes de Fourier correspondantes
Les ondes de Fourier ont la propriété que, si ϕ est une onde de Fourier,

$$\forall x, y \in [-\frac{1}{2}, \frac{1}{2}[\ \phi(x+y) = \phi(x)\phi(y)$$

Avec cette seule condition, on a

$$\exists k \in \mathbb{Z}, \ \forall x \in [-\frac{1}{2}, \frac{1}{2}[\phi(x) = e^{2i\pi kx}$$

L'entier relatif k est appelé fréquence de l'onde de Fourier ϕ.

Comme dans les autres cas, si T est un SLI et que f est une onde de Fourier de fréquence ν alors il existe une complexe $C(\nu)$ appelé gain fréquentiel ou réponse fréquentielle tel que

$$T(f) = C(\nu)f$$

Remarque 1.50. Comme pour l'espace $\{0, \ldots, N-1\}$, les fréquences sont maintenant quantifiées, mais en nombre infini. En effet, si ϕ est une onde de Fourier, on sait que $\phi(1/2) = \phi(-1/2)$ (par continuité et le fait que $1/2 = -1/2$ modulo 1). Donc $e^{2i\pi\nu} = 1$ où ν est la fréquence a priori de l'onde ϕ. Ceci oblige ν à être un entier.

La figure 1.18 montre une onde de Fourier sur $[-\frac{1}{2}, \frac{1}{2}[$. Sa fréquence est égale à 3. On remarque qu'elle parcourt trois périodes.

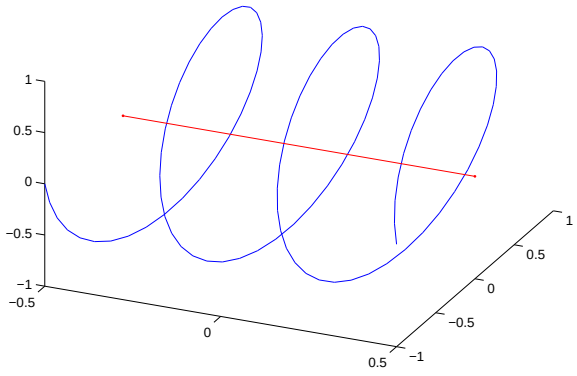


FIGURE 1.18 – **Une onde de Fourier sur $[-\frac{1}{2}, \frac{1}{2}[$ de fréquence 3. Noter comme elle parcourt exactement trois périodes le long de son domaine de définition.**

Chapitre 2

La transformation de Fourier (pour $\mathbb{Z}$ et $\{0, \ldots, N-1\}$)

Dans ce chapitre nous allons voir les principales propriétés de la transformation de Fourier. Nous avons introduit la transformation de Fourier au chapitre précédent en tant que formule qui donne la réponse (ou gain) fréquentielle en fonction de la réponse impulsionnelle d'un SLI. Comme au chapitre précédent, nous introduisons la transformation de Fourier pour les suites (signaux définis sur $\mathbb{Z}$) avant de la présenter pour tous les autres types de signaux vus dans ce cours. Le lecteur est invité à prédire au cours de sa lecture de la première partie, ce que seront les propriétés de la transformée de Fourier sur les autres espaces. Comme pour les SLI, les propriétés de la TF sont générales. Pour chaque traitement d'une transformée de Fourier, nous présentons la définition des trois espaces de fonctions : sommables, d'énergie finie et bornées. Ensuite, nous donnons les règles de multiplication et de convolution entre ces espaces (le lecteur aura compris qu'il y a beaucoup de rapports entre l'opération de convolution et la transformée de Fourier). Enfin, nous définissons la transformée de Fourier et en donnons les propriétés. La plupart du temps les démonstrations sont absentes car les théorèmes requis pour les faire ne sont pas à notre disposition [1].

2.1 Rappel des ondes de Fourier sur les différents espaces

On rappelle la formule des ondes de Fourier sur les quatre espace $\mathbb{Z}$, $[-\frac{1}{2}, \frac{1}{2}[$, $\{0, \ldots, N-1\}$ et $\mathbb{R}$ ainsi que l'espace des fréquences pour chaque cas.

1. Une onde de Fourier sur $\mathbb{Z}$ de fréquence ν a pour formule

$$n \mapsto e^{2i\pi\nu n}$$

On remarque que les fréquences ν et $\nu+m$ (m est entier) donnent lieu à la même onde de Fourier (les valeurs en tout n sont les mêmes). C'est pourquoi nous restreignons l'espace des fréquences de $\mathbb{Z}$ à $[-\frac{1}{2}, \frac{1}{2}[$.

1. Vous trouverez sur le site pédagogique un polycopié de l'UE MDI220, Hilbert Fourier, où toutes les preuves sont faites

2. Une onde de Fourier sur $[-\frac{1}{2}, \frac{1}{2}[$ de fréquence $m \in \mathbb{Z}$ a pour formule

$$\nu \mapsto e^{2i\pi m\nu}$$

L'espace des fréquences de $[-\frac{1}{2}, \frac{1}{2}[$ est $\mathbb{Z}$

3. Une onde de Fourier sur $\mathbb{R}$ de fréquence $\nu \in \mathbb{R}$ a pour formule

$$x \mapsto e^{2i\pi\nu x}$$

L'espace des fréquences de $\mathbb{R}$ est $\mathbb{R}$ lui-même et, cette fois-ci, il n'y a pas de confusion entre l'onde de fréquence ν et celle de fréquence $\nu + 1$.

4. Une onde de Fourier sur $\{0, \ldots, N-1\}$ de fréquence $\frac{k}{M}$ (où k parcourt $\{0, \ldots, N-1\}$) a pour formule

$$n \mapsto e^{2i\pi \frac{k}{N} n}$$

L'espace des fréquences de $\{0, \ldots, N-1\}$ est donc $\frac{\{0,\ldots,N-1\}}{N}$ et peut être indexé par $\{0, \ldots, N-1\}$ lui-même.

2.2 Transformée de Fourier pour les suites, ou Transformée de Fourier à temps Discret

2.2.1 Les principaux espaces de suites et les règles de calcul

Définition 2.1. Espaces de suites
On définit les trois espaces de suite principaux suivants :

1. *On note l^1 l'espace des suites* **sommables** *c'est-à-dire des suites u qui vérifient :*

$$\sum_{n\in\mathbb{Z}} |u_n| < +\infty.$$

on note

$$\|u\|_1 = \sum_{n\in\mathbb{Z}} |u_n|$$

C'est une norme sur l'espace des suites sommables.

2. *On note l^2 l'espace des suites d'***énergie finie** *c'est-à-dire des suites u qui vérifient :*

$$\sum_{n\in\mathbb{Z}} |u_n|^2 < +\infty.$$

et on note

$$\|u\|_2 = \left(\sum_{n\in\mathbb{Z}} |u_n|^2\right)^{\frac{1}{2}}$$

Il s'agit d'une norme sur l'espace des suites d'énergie finie.

3. *On note l^∞ l'espace des suites* **bornées** *c'est-à-dire des suites u qui vérifient :*

$$\exists C \in \mathbb{R}_+,\ \forall n \in \mathbb{Z},\ |u_n| \leqslant C.$$

et on note

$$\|u\|_\infty = sup_{n\in\mathbb{Z}}\{|u_n|\}$$

c'est une norme sur l'espace des suites bornées.

Et on a

$$l^1 \subset l^2 \subset l^\infty$$

Ces inclusions sont un cas particulier des suites sur $\mathbb{Z}$, elles font que certains théorèmes tels que le théorème d'inversion ont des hypothèses plus faibles que sur d'autres espaces

Démonstration.
Nous allons seulement prouver les inclusions :

1. $l^1 \subset l^2$: Soit $u \in l^1$. Soit $E \subset \mathbb{Z}$ défini par $E = \{n : |u_n| > 1\}$. L'ensemble E est forcément fini, sinon u aurait une somme infinie. Et on a

$$\sum_{n\in\mathbb{Z}} |u_n|^2 = \sum_{n\in E} |u_n|^2 + \sum_{n\notin E} |u_n|^2 \leqslant \sum_{n\in E} |u_n|^2 + \sum_{n\notin E} |u_n|$$

   La dernière inégalité vient du fait que si $|x| \leqslant 1$ alors $|x|^2 \leqslant |x|$. Or le premier terme de la dernière somme est fini, car E est fini. Le second est également fini, car u est sommable. Donc $u \in l^2$.

2. $l^2 \subset l^\infty$: Il est clair que si u n'est pas bornée alors elle a, par exemple, une infinité de termes supérieurs à 1 en module. Les $|u_n|^2$ ne pourraient donc pas être sommables.

□

Définition 2.2. rappel de l'opération de convolution
On rappelle que si u et v sont des suites, on appelle produit de convolution de u et v la suite définie par (si la somme a un sens) :

$$(u * v)_n = \sum_{m\in\mathbb{Z}} u_m v_{n-m}.$$

On renvoie le lecteur au chapitre précédent pour les propriétés élémentaires de la convolution (commutativité, linéarité et associativité). Nous donnons dans la proposition suivante les règles qui disent, suivant les espaces où se trouvent u et v, l'espace où se trouve $u * v$ et $u.v$ (produit point à point des deux suites).

Proposition 2.3. Inégalité de Hölder

1. *Si $u \in l^1$ et $v \in l^\infty$, alors $u.v \in l^1$ et*

$$\|u.v\|_1 \leqslant \|u\|_1 \|v\|_\infty$$

2. *Si $u \in l^2$ et $v \in l^2$, alors $u.v \in l^1$ et*

$$\|u.v\|_1 \leqslant \|u\|_2 \|v\|_2$$

   (cette inégalité s'appelle aussi inégalité de Cauchy-Schwartz)

Proposition 2.4. Règles de convolution
Le tableau suivant se lit comme suit, si u appartient à un espace index de ligne et v se trouve dans l'espace index de colonne, alors $u * v$ se trouve dans l'espace inscrit dans la case correspondante. Par exemple, si $u \in l^1$ et $v \in l^\infty$ alors $u * v \in l^\infty$. Si la case contient un tiret (-) alors l'opération est a priori impossible (la somme infinie peut ne pas avoir de sens). Et on a aussi à chaque fois, $\|u * v\|_\gamma \leqslant \|u\|_\alpha \|v\|_\beta$ (où les normes α, β et γ sont les normes index de colonne, ligne et norme dans la case indexée respectivement. Par exemple $\|u * v\|_\infty \leqslant \|u\|_1 \|v\|_\infty$)

$*$	l^1	l^2	l^∞
l^1	l^1	l^2	l^∞
l^2	l^2	l^∞	-
l^∞	l^∞	-	-

Remarque 2.5.

La démonstration du cas l^1 contre l^∞ se déduit simplement de l'inégalité de Hölder. Vous pouvez la faire. Le cas l^2 contre l^2 aussi découle de l'inégalité de Hölder (ou Cauchy-Schwartz). Le dernier cas est celui de l^1 contre l^1, il découle du théorème de Fubini.

Le lecteur pourra relire le théorème 1.13 à la lumière de ces règles pour constater que ce sont elles qui nous ont menés à la distinction des différents cas de SLI que comporte ce théorème.

2.2.2 La Transformée de Fourier sur $\mathbb{Z}$: Transformée de Fourier à temps Discret (TFtD)

Définition 2.6.
Les suites du type $n \mapsto e^{2i\pi\nu n}$ sont appelée **ondes de Fourier** *sur $\mathbb{Z}$ et on remarque que l'on peut restreindre ν à l'intervalle $[-\frac{1}{2}, \frac{1}{2}[$ pour toutes les décrire. ν est appelé* **fréquence** *d'une telle onde.*

Si u est une suite sommable ($u \in l^1$), on appelle Transformée de Fourier à temps Discret (TFtD en abrégé), la fonction définie sur l'intervalle $[-\frac{1}{2}, \frac{1}{2}[$ et que l'on note soit $\hat{u}$ soit $\mathcal{F}(u)$, qui vaut

$$\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[,\ \hat{u}(\nu) = \sum_{n \in \mathbb{Z}} u_n e^{-2i\pi\nu n}$$

Quand $u \in l^1$, $\hat{u}$ est une fonction continue et admet une limite en $\frac{1}{2}$ égale à sa valeur en $-\frac{1}{2}$ (i.e. elle est continue même si on la regarde comme une fonction définie sur $\mathbb{R}$ obtenue en périodisant la définition sur $[-\frac{1}{2}, \frac{1}{2}[$).

Le fait que la définition ait un sens découle du fait que u est supposée sommable. Le fait que $\hat{u}$ soit continue utilise le théorème de convergence dominée (voir polycopié disponible sur le site pédagogique). On peut aussi montrer la continuité de $\hat{u}$ on la regardant comme une série sommable de fonctions du type $\nu \mapsto u_n e^{2i\pi\nu n}$ dans l'espace des fonctions continues sur l'intervalle $[-\frac{1}{2}, \frac{1}{2}[$ qui est un espace complet.

Proposition 2.7. Propriétés de la TFtD
Dans la suite u et v sont des suites de l^1 et ν_0 un élément de $[-\frac{1}{2}, \frac{1}{2}[$ et $\varphi_n = e^{2i\pi\nu_0 n}$ est

une onde de Fourier sur $\mathbb{Z}$ de fréquence ν_0. Enfin, $m \in \mathbb{Z}$ et $\psi(x) = e^{-2i\pi mx}$ est une onde de Fourier sur $[-\frac{1}{2}, \frac{1}{2}[$ de fréquence $-m$.

1. *La TFtD de l'impulsion en m est une onde de Fourier de fréquence $-m$ sur $[-\frac{1}{2}, \frac{1}{2}[$:*

$$(\forall n \in \mathbb{Z},\ u_n = \delta_n^m) \Longrightarrow \left(\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[,\ \hat{u}(\nu) = \psi(\nu) = e^{-2i\pi m\nu} \quad \right)$$

2. *La convolution est transformée en produit :*

$$\mathcal{F}(u * v) = \hat{u}\hat{v}$$

3. *Le produit est transformé en convolution :*

$$\mathcal{F}(uv) = \hat{u} * \hat{v}$$

   (sur d'autres espaces que $\mathbb{Z}$ il faut prendre d'autres précautions sur u et v)

4. *Multiplier par une onde de Fourier de fréquence ν_0 revient à décaler la transformée de ν_0 :*

$$\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[,\ [\mathcal{F}(\varphi.u)](\nu) = \hat{u}(\nu - \nu_0)$$

   Avec la règle de tanslation des fonctions de $[-\frac{1}{2}, \frac{1}{2}[$ vue au chapitre précédent.

5. *Un décalage de la suite de m revient à multiplier la TFtD par une onde de Fourier de fréquence $-m$. On note u^m la m-translatée de u ($u_n^m = u_{n-m}$) :*

$$\mathcal{F}(u^m) = \hat{u}.\psi, \qquad \text{ou encore } \widehat{u^m}(\nu) = \hat{u}(\nu)e^{-2i\pi m\nu}$$

   (On peut voir cette propriété comme une application des propriétés 1 et 2, en remarquant que décaler une suite de m revient à la convoluer contre l'impulsion en m).

6. *Si u est réelle, alors $\hat{u}$ possède la* **symétrie hermitienne** *:*

$$(\forall n \in \mathbb{Z},\ u_n \in \mathbb{R}) \Longrightarrow \left(\ \forall \nu \in [-\frac{1}{2}, \frac{1}{2}[,\ \hat{u}(-\nu) = \overline{\hat{u}(\nu)} \quad \right)$$

   (En particulier $\hat{u}(-\frac{1}{2}) \in \mathbb{R}$)

7. *Si u est symétrique alors $\hat{u}$ aussi :*

$$(\forall n \in \mathbb{Z},\ u_{-n} = u_n) \Longrightarrow \left(\ \forall \nu \in [-\frac{1}{2}, \frac{1}{2}[,\ \hat{u}(-\nu) = \hat{u}(\nu) \quad \right)$$

8. *Si u est à la fois symétrique et réelle alors $\hat{u}$ est aussi symétrique et réelle (cela se déduit des deux propriétés précédentes).*

$$(\forall n \in \mathbb{Z},\ u_{-n} = u_n \in \mathbb{R}) \Longrightarrow \left(\ \forall \nu \in [-\frac{1}{2}, \frac{1}{2}[,\ \hat{u}(-\nu) = \hat{u}(\nu) \in \mathbb{R} \quad \right)$$

Démonstration 2.8.

1. La suite est donc nulle partout sauf pour $u_m = 1$, on écrit la définition de la TFtD et on a
$$\hat{u}(\nu) = \sum_{n\in\mathbb{Z}} u_n e^{-2i\pi\nu\ n} = u_m e^{-2i\pi\nu m} = e^{-2i\pi\nu m} = \psi(\nu)$$

2. Si u et v sont des suites de l^1, les règles de calcul nous disent que leur produit de convolution est bien l^1 (sommable) et il est donc légitime d'en considérer la TFtD. Le résultat découle du théorème de Fubini. On peut aussi voir ce résultat en décomposant u et v comme sommes d'impulsions. L'opération de convolution étant bilinéaire (en u et v) il suffit de démontrer la formule pour des impulsions (**faites-le**) et de conclure en utilisant des passages à la limite dans des espaces de dimension infinie (l^1 et l^∞ en l'occurence).

3. Contentons-nous de démontrer le résultat dans le cas où u et v sont des impulsions. Posons donc $u_n = \delta_n^{m_1}$ et $v_n = \delta_n^{m_2}$.
D'après la première propriété $\hat{u}(\nu) = e^{-2i\pi m_1\nu}$ et $\hat{v}(\nu) = e^{-2i\pi m_2\nu}$. La convolution de deux ondes de Fourier sur $[-\frac{1}{2}, \frac{1}{2}[$ est possible (les ondes de Fourier sont sommables sur $[-\frac{1}{2}, \frac{1}{2}[$ alors qu'elles ne le sont pas sur $\mathbb{Z}$) et le produit de leur convolution est
$$\int_{-\frac{1}{2}}^{\frac{1}{2}} e^{-2i\pi m_1 t} e^{-2i\pi m_2(\nu-t)} dt = e^{-2i\pi m_2\nu} \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{-2i\pi(m_1-m_2)t} dt = \begin{cases} e^{-i\pi m_2\nu} = e^{-i\pi m_2\nu} \text{ si } m_1 = m_2 \\ 0 \text{ si } m_1 \neq m_2 \end{cases}$$
Soit la fonction nulle si m_1 et m_2 sont différents et l'une des deux ondes si $m_1 = m_2$ (auquel cas les deux ondes sont les mêmes). Or, le produit $u.v$ a exactement le même comportement :
$$u.v = \begin{cases} u = v \text{ si } m_1 = m_2 \\ 0 \text{ si } m_1 \neq m_2 \end{cases}$$

4. On pose le calcul
$$\sum_{n\in\mathbb{Z}} u_n e^{2i\pi\nu_0 n} e^{-2i\pi\nu n} = \sum_{n\in\mathbb{Z}} u_n e^{-2i\pi(\nu-\nu_0)n} = \hat{u}(\nu - \nu_0)$$

5. On peut soit poser le calcul soit remarquer que u^m est le résultat de la convolution de u avec l'impulsion en m et utiliser la propriété 2 (et 1).

6. Il suffit de poser le calcul de $\overline{\hat{u}(\nu)}$
$$\overline{\hat{u}(\nu)} = \overline{\sum_{n\in\mathbb{Z}} u_n e^{-2i\pi\nu n}} = \sum_{n\in\mathbb{Z}} \overline{u_n e^{-2i\pi\nu n}} = \sum_{n\in\mathbb{Z}} u_n e^{2i\pi\nu n} = \sum_{n\in\mathbb{Z}} u_n e^{-2i\pi(-\nu)n} = \hat{u}(-\nu)$$

7. Il suffit de faire le changement de variable $m = -n$ dans la somme définissant la TFtD.

8. Si u est réelle alors (par 6)
$$\hat{u}(-\nu) = \overline{\hat{u}(\nu)}$$
Si u est symétrique alors (par 7)
$$\hat{u}(-\nu) = \hat{u}(\nu)$$

Donc $\hat{u}$ est symétrique et, en éliminant le terme de gauche des deux dernières équations :

$$\hat{u}(\nu) = \overline{\hat{u}(\nu)}$$

Ce qui montre que $\hat{u}$ est à valeurs réelles.

Proposition 2.9. SLI et transformation de Fourier
Si T est un SLI qui transforme une suite bornée en une suite bornée, alors on sait que T possède une réponse impulsionnelle sommable (voir théorème 1.13) que l'on note h.

Si u est une suite sommable (donc bornée) et $v = T(u)$ la sortie qui lui est associée par T alors on a :

1. *La réponse (ou le gain) fréquentielle du SLI T est la transformée de Fourier de h.*
2. *La suite v est sommable, car $h * v \in l^1$ d'après les règles de calcul.*
3. *Et*

$$\hat{v} = \hat{h}\hat{u}.$$

D'après les propriétés de la TFtD.

Ainsi, un SLI agit sur la transformée de Fourier de l'entrée par multiplication de celle-ci, point à point par sa réponse fréquentielle.

Démonstration.
Commençons par remarquer que les suites u, v et h ont bien toutes une transformée de Fourier à temps discret, car elles sont toutes sommables.

La seule chose qui reste à montrer est la première affirmation. Soit donc une onde de Fourier de fréquence ν, son terme général est $e^{2i\pi\nu n}$. La sortie qui lui est associée est, par définition de la réponse fréquentielle, de terme général $C(\nu)e^{2i\pi\nu n}$, où $C(\nu)$ est la réponse fréquentielle.

Or la valeur en zéro de la réponse de T à l'onde de Fourier est donnée par (convolution entre h et l'onde de Fourier) :

$$\sum_{m\in\mathbb{Z}} h_m e^{2i\pi\nu(0-m)} = \hat{h}(\nu) = C(\nu)$$

La première égalité vient de la définition de la TFtD, et la seconde de la définition de la réponse fréquentielle (la valeur en zéro de la sortie associée à une onde de Fourier). □

2.2.3 Extension à l^2 et égalité de Parseval

Nous avons défini la TFtD pour les suites sommables. Il est possible d'étendre cette définition aux suites l^2 (qui sont plus nombreuses que les suites sommables) et nous avons même l'égalité des normes entre une suite et sa transformée.

D'abord nous définissons l'espace des fonctions de $[-\frac{1}{2}, \frac{1}{2}[$ d'énergie finie qui sera le correspondant de l^2 par la TFtD.

Définition 2.10. $L^2([-\frac{1}{2}, \frac{1}{2}[)$
On dit qu'une fonction définie sur $[-\frac{1}{2}, \frac{1}{2}[$ appartient à l'espace $L^2([-\frac{1}{2}, \frac{1}{2}[)$ (espace des fonctions d'énergie finie), si :

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} |f(x)|^2 dx < +\infty$$

Et on note

$$\|f\|_2 = \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} |f(x)|^2 dx \right)^{\frac{1}{2}}$$

Il s'agit d'une norme sur l'espace vectoriel $L^2([-\frac{1}{2}, \frac{1}{2}[)$.

Théorème 2.11. Extension à l^2 et égalité de Parseval
Il existe une unique application linéaire de l^2 *vers* $L^2([-\frac{1}{2}, \frac{1}{2}[)$ *qui coïncide avec la TFtD sur les suites sommables. On note encore cette application* $\mathcal{F}$ *ou encore* $\hat{u}$ *pour signifier la TFtD de* u.

$\mathcal{F}$ est une bijection entre l^2 et $L^2([-\frac{1}{2}, \frac{1}{2}[)$.

De plus on a **l'égalité de Parseval** *:*

$$\forall u \in l^2, \ \|\hat{u}\|_2 = \|u\|_2$$

Soit plus explicitement :

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} |\hat{u}(\nu)|^2 d\nu = \sum_{n\in\mathbb{Z}} |u_n|^2$$

Démonstration.
Nous ne faisons pas la preuve de l'existence et l'unicité de l'extension de la TFtD à l^2. Disons simplement que si une suite est dans l^2 sans être sommable, on ne peut plus écrire

$$\sum_{n\in\mathbb{Z}} u_n e^{-2i\pi\nu n},$$

car cette somme n'a pas de sens *a priori*.

Par contre la suite de fonctions (indexée par N)

$$\nu \mapsto \sum_{n=-N}^{N} u_n e^{-2i\pi\nu n}$$

converge dans l'espace $L^2([-\frac{1}{2}, \frac{1}{2}[)$. C'est cette limite dans l'espace complet $L^2([-\frac{1}{2}, \frac{1}{2}[)$ que l'on prend comme TFtD d'une suite l^2.

Le coeur de la démonstration de ce théorème, que nous ne faisons pas entièrement ici, est de prouver l'égalité de Parseval. Nous allons prouver l'égalité de Parseval pour les suites à support fini.

Soit u une suite à support fini. On note e_m ($m \in \mathbb{Z}$) les ondes de Fourier sur $[-\frac{1}{2}, \frac{1}{2}[$

$$e_m(\nu) = e^{2i\pi m\nu}$$

Comme u est à support fini, sa transformée de Fourier à temps discret est égale à une combinaison finie des e_m

$$\hat{u} = \sum_{m\in\mathbb{Z}} u_m e_{-m} \text{ (somme finie par hypothèse)}$$

Or il est facile de montrer que les e_m sont orthonormés, c'est à dire

$$\|e_m\|_2 = 1$$

et

$$\forall m, n \in \mathbb{Z},\ (\quad m \neq n \quad) \Longrightarrow \int_{-\frac{1}{2}}^{\frac{1}{2}} e_m(t)\overline{e_n(t)}dt = 0$$

De cela, et comme on est dans le cas de sommes finies, on sait déjà par les propriétés des espaces hermitiens que

$$\|u\|_2 = \|\hat{u}\|_2$$

Après cela, ce qui manque pour la démonstration, sont des raisonnements de convergence dans des espaces de dimension infinie. □

Remarque 2.12.
Les propriétés 2 à 8 de la proposition 2.7 ont encore vraies en prenant u et v dans l^2, à un détail près : dans la propriété 2, il faut prendre $u \in l^2$ et $v \in l^1$, sinon la convolution de u et v n'est *a priori* ni dans l^1 ni dans l^2.

2.2.4 Théorème d'inversion

Le théorème d'inversion nous dit comment revenir de la transformée de Fourier à la suite d'origine. Vous remarquerez que la formule inverse de la transformée de Fourier à temps discret ressemble à une transformation de Fourier.

Théorème 2.13. Inversion de la TFtd
Si $u \in l^2$ est une suite d'énergie finie et $\hat{u}$ sa TFtD alors on a

$$\forall n \in \mathbb{Z},\ u_n = \int_{-\frac{1}{2}}^{\frac{1}{2}} \hat{u}(\nu)e^{2i\pi n\nu}d\nu$$

Ce théorème s'applique aussi si, $u \in l^1$ car $l^1 \subset l^2$. Sur d'autres espaces, il a des hypothèses plus restrictives.

Démonstration.
Nous ne faisons la preuve que dans le cas de suites à support fini. Soit donc u à support fini, on reprend la notation e_m pour les ondes de Fourier sur $[-\frac{1}{2}, \frac{1}{2}[$ et on rappelle que

$$\hat{u} = \sum_{m\in\mathbb{Z}} u_m e_{-m}$$

Cette somme est finie. Fixons un entier n et calculons

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \hat{u}(\nu)e^{2i\pi n\nu} = \sum_{m\in\mathbb{Z}} u_m \int_{-\frac{1}{2}}^{\frac{1}{2}} e_{-m}(\nu)e_n(\nu)d\nu = \sum_{m\in\mathbb{Z}} u_m \int_{-\frac{1}{2}}^{\frac{1}{2}} \overline{e_m(\nu)}e_n(\nu)d\nu = u_n$$

La dernière égalité vient du fait que les e_m sont orthonormés. □

Remarque 2.14. **Équivalence entre réponse impulsionnelle et réponse fréquentielle**
On a vu que la réponse fréquentielle d'un SLI est la TFtD de sa réponse impulsionnelle. Le théorème d'inversion nous permet de retrouver la réponse impulsionnelle à partir de sa réponse fréquentielle. Ainsi, pour définir un SLI, il suffit de donner soit sa réponse impulsionnelle (ce que nous savions déjà) ou sa réponse fréquentielle (qui permet de retrouver la réponse impulsionnelle grâce au théorème d'inversion).

2.2.5 Décroissance à l'infini et régularité

Nous savons déjà que si u est sommable alors $\hat{u}$ est continue. La sommabilité est une forme de décroissance à l'infini (pour les suites, cela implique même que u_n tende vers 0 à l'infini). Le théorème suivant dit que plus la suite u décroît rapidement à l'infini, plus sa TFtD est régulière.

Théorème 2.15. Décroissance à l'infini et régularité de la TFtD
Soit $k \geqslant 0$ un entier, on a :

$$\left(\sum_{n\in\mathbb{Z}} |n|^k |u_n| < \infty\right) \Longrightarrow \left(\hat{u} \text{ est } k \text{ fois continuement dérivable, soit } \hat{u} \in \mathcal{C}^k([-\frac{1}{2}, \frac{1}{2}[)\right)$$

et si on note v^k la suite de terme général

$$v_n^k = (-2i\pi n)^k u_n$$

la TFtD de v^k est

$$\mathcal{F}(v^k) = \mathcal{F}(u)^{(k)} \text{ (la dérivée } k\text{-ième de } \mathcal{F}(u) \text{)}$$

Démonstration.
Le cas $k = 0$ correspond simplement à la continuité de $\hat{u}$. Il est déjà connu.

On démontre le théorème pour les suites à support fini. La dérivée k-ième de $\hat{u}$ est donnée par (on dérive par rapport à ν, la variable de la fonction $\hat{u}$)

$$\left(\sum_{n\in\mathbb{Z}} u_n e^{-2i\pi\nu n}\right)^{(k)} = \left(\sum_{n\in\mathbb{Z}} (-2i\pi n) u_n e^{-2i\pi\nu n}\right)^{(k-1)} = \sum_{n\in\mathbb{Z}} (-2i\pi n)^k u_n e^{-2i\pi\nu n}$$

Les dérivations sous le signe somme sont légitimes, car la somme est supposée finie.

On a donc bien démontré que la dérivée k-ième de $\hat{u}$ (ou $\mathcal{F}(u)$) est la TFtD de la suite v^k. □

2.3 Transformée de Fourier Discrète ou TFD

La Transformée de Fourier discrète est la transformation de Fourier pour les signaux définis sur un ensemble $\{0, \ldots, N-1\}$. Les suites définies sur cet ensemble sont toutes sommables, bornées et d'énergie finie (car elles sont à support fini).

Définition 2.16. Transformée de Fourier Discrète
Si u est une suite finie définie sur $\{0, \ldots, N-1\}$, on note $\hat{u}$ et parfois U sa Transformée de Fourier Discrète (TFD en abrégé) définie, elle aussi sur $\{0, \ldots, N-1\}$, par

$$\forall k \in \{0, \ldots, N-1\},\ \hat{u}(k) = \sum_{n\in\{0,\ldots,N-1\}} u_n e^{-2i\pi\frac{k}{N}n}$$

(on pourra noter $\hat{u}_k$ au lieu $\hat{u}(k)$ pour signifier que $\hat{u}$ est aussi une suite).

Théorème 2.17. Théorème d'inversion et interprétation de la TFD comme décomposition sur une base
Si u est une suite définie sur $\{0,\dots,N-1\}$ et $\hat{u}$ sa TFD on a :

$$\forall n \in \{0,\dots,N-1\},\ u_n = \frac{1}{N} \sum_{k\in\{0,\dots,N-1\}} \hat{u}_k e^{2i\pi\frac{k}{N}n}$$

Ceci peut se réécrire sous la forme

$$u = \sum_{k\in\{0,\dots,N-1\}} \frac{1}{N}\hat{u}_k \mathbf{w}^k$$

où $\mathbf{w}^k$ est l'onde de Fourier sur $\{0,\dots,N-1\}$ de fréquence k/N. En d'autres termes, les $\frac{\hat{u}(k)}{N}$ sont les coefficients de la décomposition de u sur la base des ondes de Fourier.

Démonstration.
Fixons n et développons

$$\frac{1}{N} \sum_{k\in\{0,\dots,N-1\}} \hat{u}_k e^{2i\pi\frac{k}{N}n} = \frac{1}{N} \sum_{k\in\{0,\dots,N-1\}} \left(\sum_{m\in\{0,\dots,N-1\}} u_m e^{-2i\pi\frac{k}{N}m} \right) e^{2i\pi\frac{k}{N}n}$$

$$= \frac{1}{N}\sum_{k,m} u_m e^{2i\pi\frac{k}{N}(n-m)} = \frac{1}{N}\sum_m u_m \left(\sum_k e^{2i\pi\frac{k}{N}(n-m)} \right) = u_n$$

La dernière égalité ayant lieu, car la somme sur k est nulle sauf si $n = m$ auquel cas elle vaut N. □

La TFD a les propriétés suivantes, que vous pouvez comparer à la proposition 2.7. Mis à part que la TFD est indexée par k quand on veut signifier la fréquence k/N, rien ne change.

Proposition 2.18. propriétés de la TFD
Dans la suite u et v sont des suites définies sur $\{0,\dots,N-1\}$ et k_0 un élément de $\{0,\dots,N-1\}$ et $\varphi_n = e^{2i\pi\frac{k_0}{N}n}$ est une onde de Fourier sur $\{0,\dots,N-1\}$ de fréquence k_0/N. Enfin, $m \in \{0,\dots,N-1\}$ et $\psi_n = e^{-2i\pi\frac{m}{N}n}$ est une onde de Fourier sur $\{0,\dots,N-1\}$ de fréquence $-m/N$.

1. *La TFD de l'impulsion en m est une onde de Fourier de fréquence $-m/N$ sur $\{0,\dots,N-1\}$:*

   $$(\forall n \in \{0,\dots,N-1\},\ u_n = \delta_n^m) \Longrightarrow \big(\forall k \in \{0,\dots,N-1\},\ \hat{u}(k) = \psi_k = e^{-2i\pi\frac{m}{N}k} \quad \big)$$

2. *La convolution est transformée en produit :*

   $$\mathcal{F}(u * v) = \hat{u}\hat{v}$$

3. *Le produit est transformé en convolution (à un facteur de normalisation près) :*

   $$\mathcal{F}(uv) = \frac{1}{N}\hat{u} * \hat{v}$$

4. *Multiplier par une onde de Fourier de fréquence k_0/N revient à décaler la transformée de k_0 :*

$$\forall k \in \{0,\dots,N-1\},\ [\mathcal{F}(\varphi.u)](k) = \hat{u}(k-k_0)$$

Avec la règle de tanslation des fonctions de $\{0,\dots,N-1\}$ vue au chapitre précédent.

5. *Un décalage de la suite de m revient à multiplier la TFD par une onde de Fourier de fréquence $-m$. On note u^m la m-translatée de u ($u^m_n = u_{n-m}$) :*

$$\mathcal{F}(u^m) = \hat{u}.\psi$$

(On peut voir cette propriété comme une application des propriétés 1 et 2, en remarquant que décaler une suite de m revient à la convoluer contre l'impulsion en m).

6. *Si u est réelle, alors $\hat{u}$ possède la* **symétrie hermitienne** *:*

$$(\forall n \in \{0,\dots,N-1\},\ u_n \in \mathbb{R}) \Longrightarrow \Big(\ \forall k \in \{0,\dots,N-1\},\ \hat{u}(-k) = \overline{\hat{u}(k)} \quad \Big)$$

(En particulier, si N est pair, $\hat{u}(-\frac{N}{2}) \in \mathbb{R}$) Ici $-k$ signifie l'inverse module N, par exemple l'inverse de 2 est $-2 = N-2$ modulo N.

7. *Si u est symétrique alors $\hat{u}$ aussi :*

$$(\forall n \in \{0,\dots,N-1\},\ u_{-n} = u_n) \Longrightarrow (\ \forall k \in \{0,\dots,N-1\},\ \hat{u}(-k) = \hat{u}(k) \quad)$$

8. *Si u est à la fois symétrique et réelle alors $\hat{u}$ est aussi symétrique et réelle (cela se déduit des deux propriétés précédentes).*

$$(\forall n \in \{0,\dots,N-1\},\ u_{-n} = u_n \in \mathbb{R}) \Longrightarrow (\ \forall k \in \{0,\dots,N-1\},\ \hat{u}(-k) = \hat{u}(k) \in \mathbb{R} \quad)$$

Proposition 2.19. Égalité de Parseval (à normalisation près)
On a l'égalité suivante

$$\sum_{n\in\{0,\dots,N-1\}} |u_n|^2 = \frac{1}{N} \sum_{k\in\{0,\dots,N-1\}} |\hat{u}_k|^2$$

où u est une suite sur $\{0,\dots,N-1\}$ et $\hat{u}$ sa TFD.

En termes de normes hermitienne, cela s'écrit

$$\|u\|_2^2 = \frac{1}{N}\|\hat{u}\|_2^2$$

et

$$\|u\|_2 = \frac{1}{\sqrt{N}}\|\hat{u}\|_2$$

Démonstration.
Comme dans le cas de la TFtD on constate que les ondes de Fourier sur $\{0,\dots,N-1\}$ sont orthogonales les unes aux autres. Or les $\hat{u}_k$ sont les coefficients de la décomposition de u sur la base des $\frac{1}{N}\mathbf{w}^k$ qui ont pour norme $1/\sqrt{N}$. Donc les $\frac{1}{\sqrt{N}}\hat{u}_k$ sont les coefficients de la décomposition de u sur la base $\frac{1}{\sqrt{N}}\mathbf{w}^k$ qui est orthonormée. On a donc

$$\sum_{n\in\{0,\dots,N-1\}} |u_n|^2 = \sum_{k\in\{0,\dots,N-1\}} \left|\frac{1}{\sqrt{N}}\hat{u}_k\right|^2 = \frac{1}{N} \sum_{k\in\{0,\dots,N-1\}} |\hat{u}_k|^2$$

□

2.4 Lien entre TFD et TFtD

La TFD est la seule transformée de Fourier calculable sur ordinateur. La plupart des signaux naturels sont définis sur un espace à la fois infini et continu ($\mathbb{R}$ pour le son, par exemple). La TFD s'intéresse à des signaux définis sur un espace fini et discret. Le passage du continu au discret sera vu au chapitre échantillonnage. Dans cette partie nous allons voir comment une TFD peut permettre d'analyser un signal défini sur $\mathbb{Z}$, réalisant ainsi un passage de l'infini au fini.

Évidemment, une TFD ne peut capturer toutes les caractéristiques d'un signal d'extension infini, il nous faut restreindre notre étude à des cas particuliers. Nous verrons :

1. Cas d'une suite à support fini.
2. Détermination de la fréquence d'une onde.
3. Séparation de la somme de deux ondes et l'outil "fenêtrage".

2.4.1 Cas d'une suite à support fini

Soit une suite u définie sur $\mathbb{Z}$ à support fini. Sans perte de généralité, quitte à la translater, on peut supposer qu'il existe un entier N tel que :

$$\forall n \in \mathbb{Z},\ u_n = 0 \text{ si } n < 0 \text{ ou } n \geqslant N$$

En d'autres termes nous supposons que u est nulle hors de l'ensemble $\{0, \dots, N-1\}$.

Pour tout entier $M \geqslant N$ on considère la suite finie v définie sur $\{0, \dots, M-1\}$ par :

$$\forall n \in \{0, \dots, M-1\},\ v_n = u_n$$

La suite finie v est simplement la restriction de u à l'ensemble $\{0, \dots, M-1\}$. On appelle parfois v la suite zéro-padding à l'ordre M de u. On ajoute des zéros à la suite des échantillons non nuls de u pour obtenir une suite de taille M.

Posons la définition de la TFD de v

$$\forall k \in \{0, \dots, M-1\},\ \hat{v}(k) = \sum_{n \in \{0,\dots,M-1\}} v_n e^{-2i\pi \frac{k}{M} n} = \hat{u}(\frac{k}{M})$$

où $\hat{u}$ est la TFtD de u. la dernière égalité provient du fait que u est nulle hors de $\{0, \dots, N-1\}$ (et donc hors de $\{0, \dots, M-1\}$).

Remarque 2.20. **Une TFD peut capturer toute l'information d'une suite à support fini**
Le théorème d'inversion pour la TFD nous dit que nous pouvons retrouver les $v_0, \dots, v_{M-1}$ à partir de la TFD de v. En particulier, on peut retrouver tous les échantillons non nuls de u à partir de la TFD de v dès que $M \geqslant N$. Ainsi, la TFD capture toute l'information de u, si u est a support fini.

Définition 2.21. TFD d'ordre arbitraire
Ainsi définie $\hat{v}$ est parfois appelée **TFD d'ordre** M *de la suite* $u_0, \dots, u_{N-1}$.

$\hat{v}(k)$ est l'échantillonnage de la TFtD de u aux points k/M lorsque k parcourt $\{0, \dots, M-1\}$. *C'est-à-dire*

$$\forall k \in \{0, \dots, M-1\},\ \hat{v}(k) = \hat{u}\left(\frac{k}{M}\right)$$

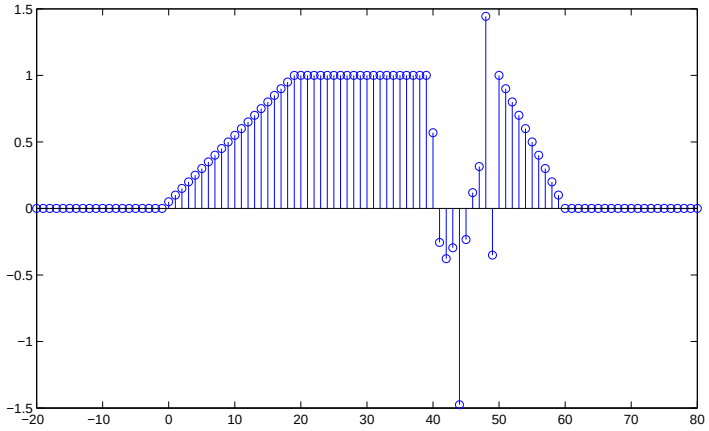


FIGURE 2.1 – **Un signal dont seuls 60 échantillons sont non nuls.**

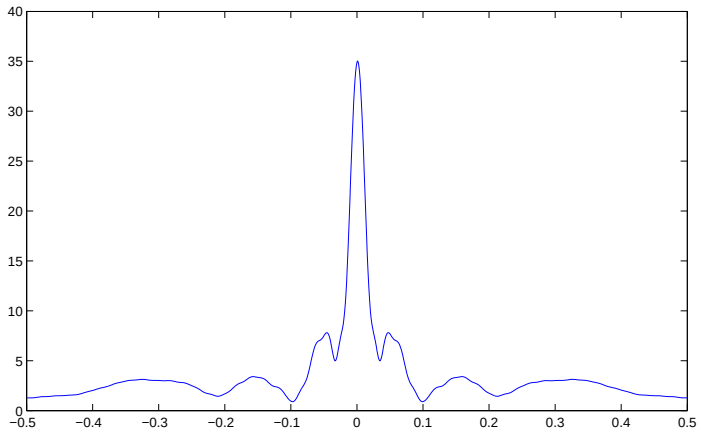


FIGURE 2.2 – **(module de la) TFtD du signal précédent. Comme le signal est à valeurs réelles sa TFtD est à symétrie hérmitienne, et comme on trace le module, on a une symétrie par rapport l'axe des ordonnées.**

N'oubliez pas que les ondes de Fourier sur $\mathbb{Z}$ *ont une fréquence qui est définie modulo 1 (L'onde de fréquence* ν *vaut la même chose en tout point de* $\mathbb{Z}$ *que l'onde de fréquence* $1+\nu$*). Ainsi, la TFD de* v *en* $M-l$ *vaut*

$$\hat{v}(M-l) = \hat{u}(\frac{M-l}{M}) = \hat{u}(1-\frac{l}{M}) = \hat{u}(-\frac{l}{M})$$

Les figures 2.1 à 2.5 montrent un signal à support fini, sa TFtD et comment la TFD en est un échantillonnage ainsi que la manière d'indexer la TFD pour la représenter dans le même repère que la TFtD.

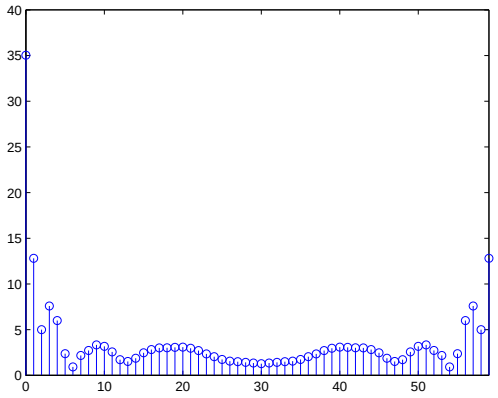


FIGURE 2.3 – **(module de la) TFD d'ordre 60 du signal précédent. Elle est indexée par $k \in \{0, \ldots, 60-1\}$. Remarquez comme les valeurs de k supérieures à $M/2 = 30$ correspondent en fait à des fréquences ν négatives.**

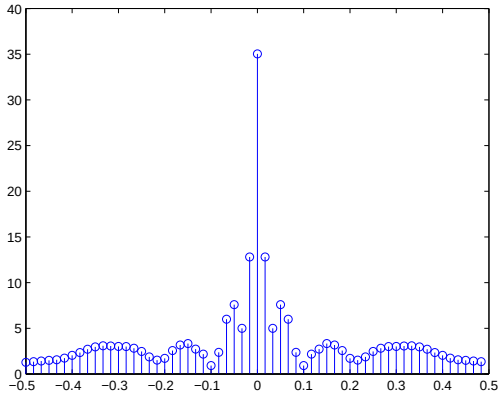


FIGURE 2.4 – **Même figure qu'à la figure 2.3 sauf que l'on a indexé la TFD par la fréquence k/M au lieu de k (et si $k/M \geqslant 1/2$ on a retranché 1 pour revenir dans $[-\frac{1}{2}, \frac{1}{2}[$).**

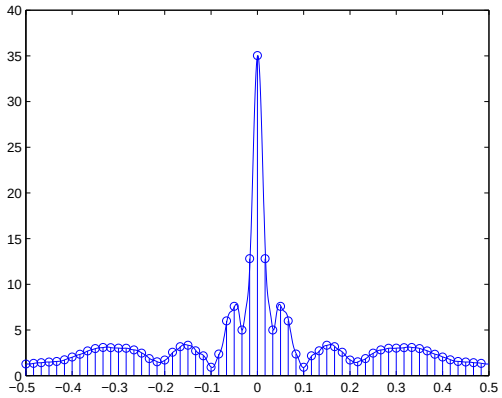


FIGURE 2.5 – **Superposition de la TFD et de la TFtD d'un même signal de support fini.**

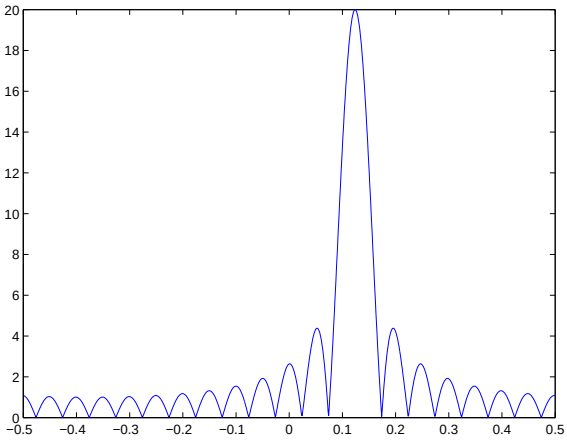


FIGURE 2.6 – **La TFtD d'une onde de fréquence 0,123 tronquée à 20 échantillons. La largeur des lobes est 1/20, sauf le lobe principal qui est de largeur 1/10.**

2.4.2 Détermination de la fréquence d'une onde grâce à une TFD

On se donne une onde $u_n = e^{2i\pi\nu_0 n}$. On voudrait déterminer la fréquence ν_0 on ne se donnant le droit que d'observer les échantillons $u_0, \ldots, u_N$. Pourquoi une telle contrainte ? Dans un signal, musical par exemple, le contenu fréquentiel évolue au cours du temps. À chaque changement de note, le signal contient des ondes de fréquences différentes. Si l'on veut, par exemple, transcrire un morceau de musique en notes, on ne peut pas se permettre une observation sur une trop longue période, car cela aurait pour effet de mélanger entre elles différentes notes. La même chose vaut pour l'analyse d'un signal de parole où l'on risque la confusion entre différents phonèmes.

On note u^T (pour "u Tronquée") la suite définie sur $\mathbb{Z}$ égale à u sur $\{0, \ldots, N-1\}$ et nulle ailleurs. Ce sont les seules valeurs que nous nous donnons le droit d'utiliser pour déterminer ν_0. Sa TFtD est

$$\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[,\ \mathcal{F}(u^T)(\nu) = e^{-i\pi.(N-1)(\nu-\nu_0)} \frac{sin(N\pi(\nu-\nu_0))}{sin(\pi(\nu-\nu_0))}$$

Le module de $\mathcal{F}(u^T)$ et donné à la figure 2.6.

Si on calcule une TFD d'ordre $M \geqslant N$, on sait que l'on va échantillonner la TFtD de u^T aux point k/M. Deux TFD d'ordres différents sont données aux figures 2.7 et 2.8.

Une démonstration fastidieuse montrerait que la valeur k/M qui se rapproche le plus de ν_0 est celle pour laquelle la TFD est maximale en module.

Proposition 2.22. Précision de la mesure d'une fréquence
Avec une TFD d'ordre M on peut connaître la fréquence de l'onde ν_0 avec une précision d'au moins $1/M$.

En effet, quelque soit ν_0 il existe au moins un k pour lequel k/M s'approche à moins $1/M$ de ν_0.

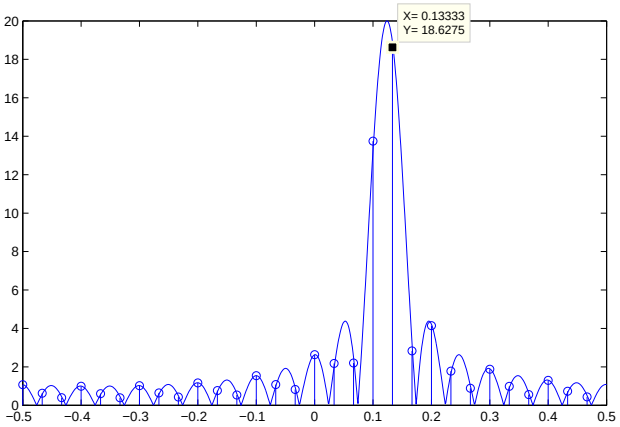


FIGURE 2.7 – **La TFD d'ordre 30 de l'onde tronquée superposée à la TFtD. Le maximum de la TFD est atteint pour $k = 4$, soit une fréquence de $4/30 = 0,1333$ et une erreur d'estimation de 0,01**

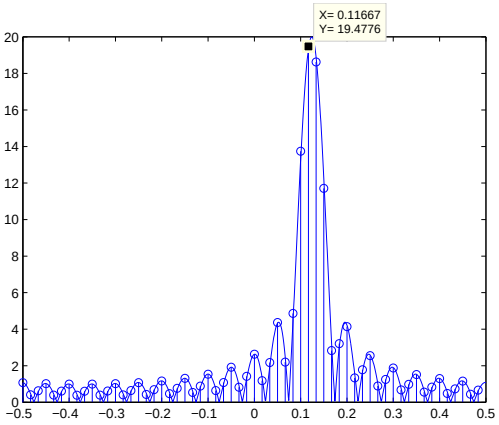


FIGURE 2.8 – **La TFD d'ordre 60 de l'onde tronquée superposée à la TFtD. Le maximum de la TFD est atteint pour $k = 7$, soit une fréquence de $7/60 = 0,11667$ et une erreur d'estimation de 0,0063**

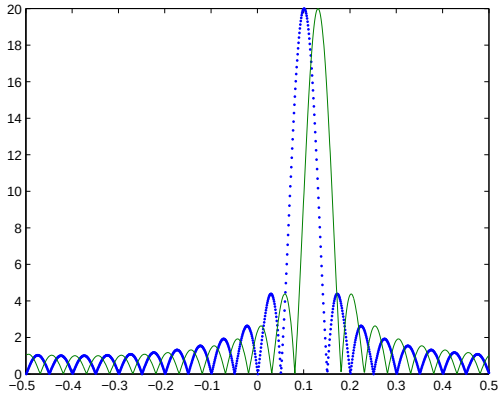


FIGURE 2.9 – **On a utilisé seulement $N = 20$ échantillons pour tracer la TFtD de deux ondes (l'une en pointillés, l'autre en trait plein) de même module et de fréquences 0,1 et 0,13.**

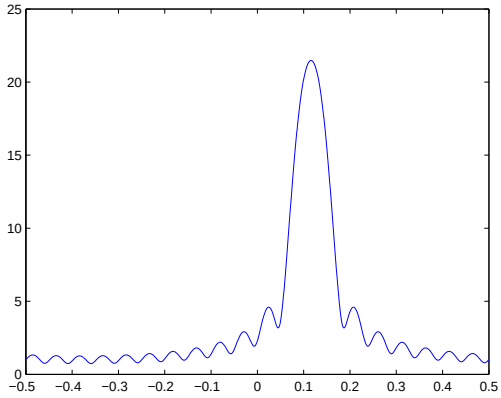


FIGURE 2.10 – **TFtD de la somme des deux ondes tronquées à 30 échantillons (figure 2.9). On ne peut pas distinguer qu'il y avait deux ondes dans le signal.**

2.4.3 Séparation de deux ondes et fenêtrage

Cette fois-ci on possède un signal plus complexe qui est la somme de deux ondes de Fourier sur $\mathbb{Z}$

$$u_n = A_0 e^{2i\pi\nu_0 n} + A_1 e^{2i\pi\nu_1 n}$$

Les inconnus ici, sont les amplitudes A_0 et A_1 ainsi que les fréquences ν_0 et ν_1. Encore une fois on ne se donne le droit que d'observer N échantillons, et on note u^T la suite ainsi tronquée.

Le problème de la résolution fréquentielle

Les graphiques de 2.9 à 2.12 illustrent le problème de la résolution fréquentielle en calculant la TFtD de u^T pour différents N.

On constate qu'il faut au moins avoir $|\nu_0 - \nu_1| > 1/N$ pour pouvoir distinguer deux pics sur la TFtD. Sinon, les deux pics se confondent en un seul et il sera impossible de distinguer ν_0 et ν_1.

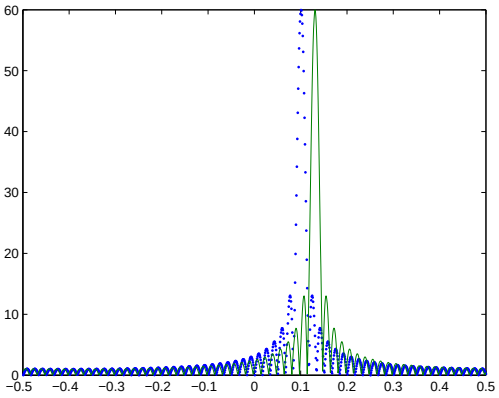


FIGURE 2.11 – **On a utilisé seulement $N = 20$ échantillons pour tracer la TFtD de deux ondes (l'une en pointillés, l'autre en trait plein) de même module et de fréquences 0,1 et 0,13.**

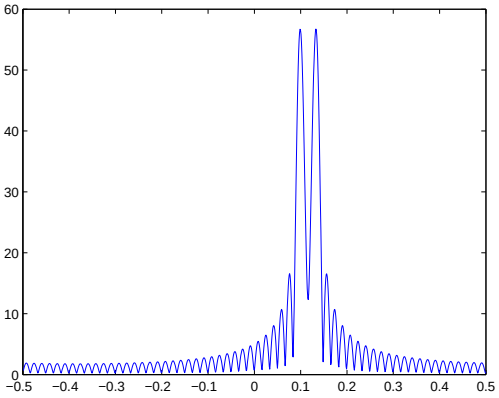


FIGURE 2.12 – **TFtD de la somme des deux ondes tronquées à 60 échantillons (figure 2.11). Cette fois on distingue bien les deux ondes. Il a fallu prendre un nombre d'échantillons N supérieur à 1/(0.13-0.1)=33.3 pour arriver à distinguer les deux.**

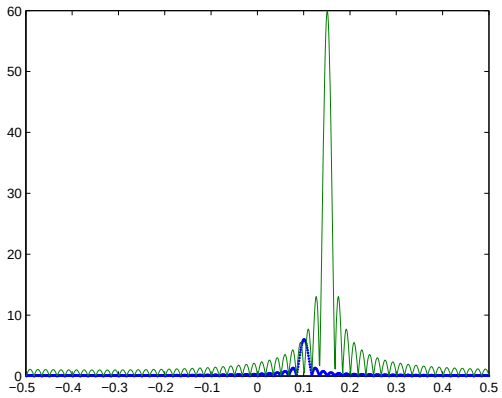


FIGURE 2.13 – **TFtD de deux ondes tronquées dont l'une a une amplitude 10 fois plus petite que l'autre. La plus petite des deux est cachée par les lobes secondaires de la TFtD de l'autre.**

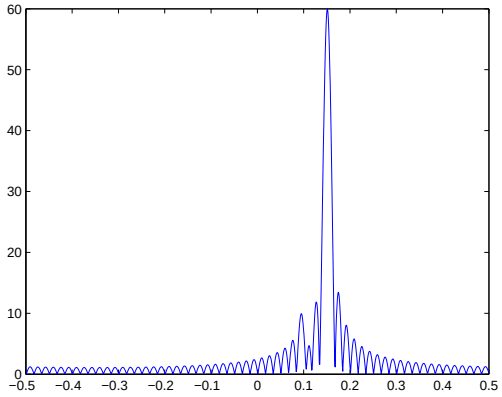


FIGURE 2.14 – **TFtD de la somme des deux ondes tronquées. Il est difficile de déceler la présence de l'onde de faible amplitude.**

On dit que $\frac{1}{N}$ est la **résolution fréquentielle**. Il faut augmenter N pour pouvoir séparer des fréquences proches l'une de l'autre.

Le problème du masquage du à une grande différence d'amplitude

Les graphiques 2.13 et 2.14 montrent une situation où A_1 est beaucoup plus grand que A_0. On constate que les lobes secondaires de la TFtD de l'onde (tronquée) ν_1 cachent jusqu'au lobe principal de l'onde tronquée de fréquence ν_0. Cela est du au fait que la troncature choisie est trop brutale. Pour obtenir u^T nous avons multiplié u par une fenêtre créneau que l'on va noter c :

$$c_n = \begin{cases} 1 \text{ si } 0 \leqslant n < N \\ 0 \text{ sinon} \end{cases}$$

et on a fait $u^T = u.c$. Comme vu plus haut la TFtD de u^T est la convolution de la

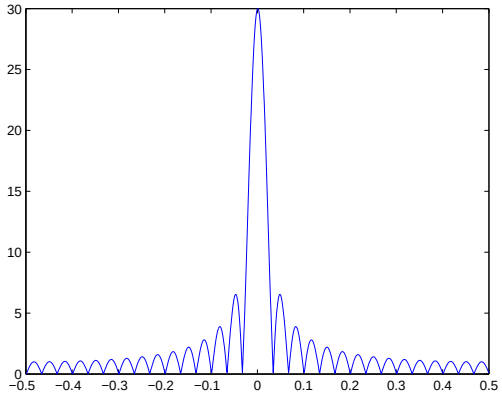


FIGURE 2.15 – **TFtD d'un créneau de taille 30.**

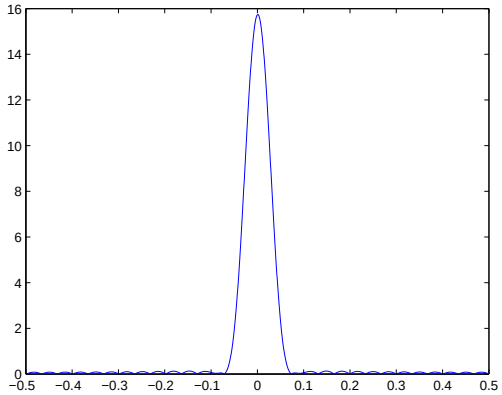


FIGURE 2.16 – **TFtD d'une fenêtre de hamming de taille $N = 30$.**

TFtD de u avec celle de c. Le graphique 2.15 montre le module de la TFtD de c. Si l'on trouve une fenêtre dont la TFtD présente des lobes secondaires moins proéminents on peut espérer résoudre le problème que pose la séparation des deux ondes de notre mélange.

Une fenêtre proposée est la fenêtre de hamming définit par

$$h_n = \begin{cases} 0.54 - 0.46cos(2\pi\frac{n}{N-1}) \text{ si } 0 \leqslant n < N \\ 0 \text{ sinon} \end{cases}$$

La figure 2.16 montre la TFtD de la fenêtre de hamming. Par rapport à celle de c (créneau), on constate deux choses

1. Le lobe central est plus étalé : ceci implique une perte de résolution fréquentielle. On passe d'une résolution de l'ordre de $1/N$ à une résolution de l'ordre de $2/N$.
2. Les lobes secondaires sont bien moins hauts que le lobe central, cela permet de distinguer deux ondes dont les amplitudes diffèrent grandement.

Les figures 2.17 et 2.18 montrent comment la multiplication par une fenêtre de hamming plutôt qu'une troncature brutale permet de distinguer une onde de faible amplitude.

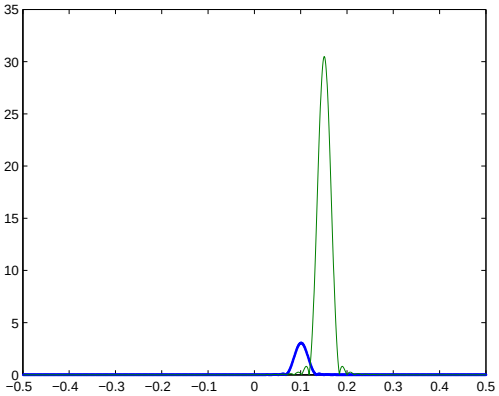


FIGURE 2.17 – **Même figure que 2.13 mais en ayant multiplié les signaux par une fenêtre de hamming plutôt que de les tronquer brutalement. Cette fois l'onde de faible amplitude est bien au dessus des lobes secondaires de l'onde de forte amplitude.**

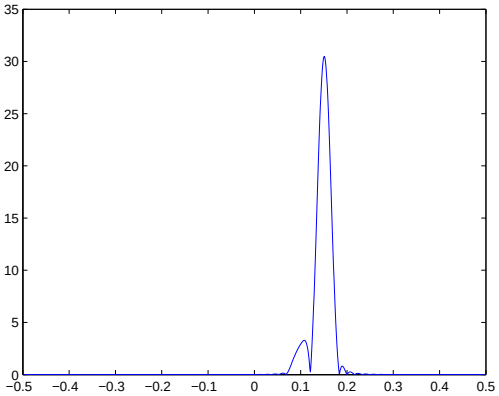


FIGURE 2.18 – **Même figure que 2.14 mais en ayant multiplié le signal contre une fenêtre de hamming plutôt que de la tronquer brutalement. Cette fois-ci on constate bien un deuxième lobe qui ne peut être confondu avec le lobe secondaire engendré par l'onde prédominante, car ces lobes secondaires sont bien plus faibles d'après la figure 2.16**

2.4.4 Conclusion sur le rapport entre TFD et TFtD

La TFD est un moyen d'approcher une TFtD. Pour obtenir une bonne résolution fréquentielle il faut observer un plus grand nombre d'échantillons du signal d'origine. Pour supprimer les effets de bord, il faut multiplier le signal par une fenêtre qui fait perdre un peu en résolution fréquentielle mais permet de distinguer plus de composantes fréquentielles. Le choix du nombre de points de la TFD sert à augmenter la précision d'échantillonnage de la TFtD de la partie du signal choisie.

2.5 TFCT

La Transformée de Fourier à Court Terme (TFCT) n'est pas à proprement parler une transformation de Fourier. Elle n'en a pas les propriétés algébriques remarquables, c'est pourtant un outil essentiel en traitement du signal. Cet outil est basé sur la constatation déjà faite plus haut que le contenu fréquentiel d'un signal peut évoluer au cours du temps. Elle se définit naïvement comme une analyse locale des composantes fréquentielles du signal. Plus précisément, pour chaque instant n, on extrait un certain nombre d'échantillons du signal étudié autour du point n que l'on étudie par les moyens vus ci-dessus (fenêtrage et TFD d'ordres arbitraires).

Définition 2.23. *Si u est une suite définie sur $\mathbb{Z}$. On fixe une fenêtre $w_0, \dots, w_n$ de taille N et on choisit un entier $M \geqslant N$. La Transformée de Fourier à Court Terme de u de fenêtre w et de précision $1/M$ est une fonction, que l'on note U définie sur $\mathbb{Z} \times \frac{\{0,\dots,M-1\}}{M}$ par*

$$\forall (n,k) \in \mathbb{Z} \times \{0,\dots,M-1\},\ U\left(n, \frac{k}{M}\right) = \sum_{m\in\mathbb{Z}} u_m w_{n-m} e^{-2i\pi \frac{k}{M} m}$$

On peut aussi considérer U comme une fonction définie sur $\mathbb{Z}\times[-\frac{1}{2}, \frac{1}{2}[$ que l'on échantiollonnera aussi finement que l'on veut suivant la seconde variable en augmentant la valeur de M (ordre de la TFD)

$$\forall (n,\nu) \in \mathbb{Z} \times [-\frac{1}{2}, \frac{1}{2}[,\ U(n,\nu) = \sum_{m\in\mathbb{Z}} u_m w_{m-n} e^{-2i\pi\nu m}$$

On distingue cette notation de la notation U pour la TFD par le fait qu'elle possède, ici, deux variables.

Interprétation

Pour n fixé : On remarque que pour un entier n fixé, la fonction $\nu \mapsto U(n,\nu)$ est la TFtD de la suite $l \mapsto u_l w_{l-n}$, c'est-à-dire la suite u multipliée par la n-translatée de la fenêtre w. Il s'agit bien de ce que nous avions annoncé, atour de chaque n, on extrait un morceau de signal dont on calcule la TFtD (par le moyen d'une TFD aussi fine que voulu).

Pour ν fixé : Pour une fréquence ν fixée avec n variable, on a :

$$U(n,\nu) = \sum_{m\in\mathbb{Z}} u_m w_{m-n} e^{-2i\pi\nu m} = e^{-2i\pi\nu n} \sum_m u_m \gamma_{n-m}$$

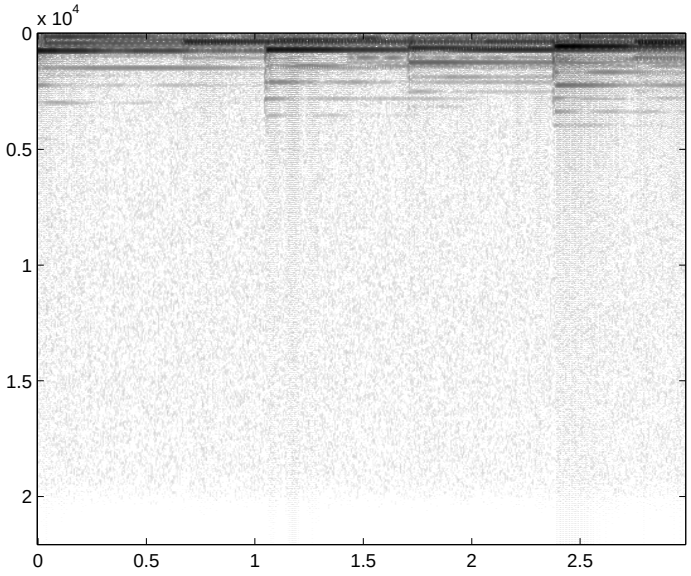


FIGURE 2.19 – **Spectrogramme d'un morceau de piano. On voit se succéder les notes. Chaque note est caractérisée par l'apparition de raies qui s'affaiblissent à mesure que le son s'atténue. (l'échelle des fréquences est du haut vers le bas et en 10000Hz d'unité.**

où γ est la suite définie par

$$\gamma_l = w_{-l} e^{2i\pi\nu l}$$

Le module de U est alors

$$|U(n,\nu)| = |(u * \gamma)_n|$$

Ainsi le module de U est celui de la convolution de la symétrique de w multipliée par une onde de fréquence ν. Cela signifie qu'à ν fixé, le module de U reflète à quel point la fréquence ν est présente dans le signal autour du point n. En effet, la TFtD de γ est centrée autour de la fréquence ν (la fenêtre w, si c'est une fenêtre de Hamming par exemple, a son spectre centré en zéro).

2.5.1 Spectrogramme et représentation graphique

Le spectrogramme est le module au carré de la TFCT ($(n,\nu) \mapsto |U(n,\nu)|^2$). On le visualise comme une image, les deux axes sont ceux des variables n et ν, on représente la valeur du spectrogramme soit en gris, suivant la valeur (sombre pour grand et clair pour faible). Ou alors en couleurs (bleu pour faible et rouge pour élevé).

Parfois, on trace le logarithme de la TFCT et non de manière linéaire, car certaines fréquences sont tellement présentes qu'elles couvrent toutes les autres si l'affichage était linéaire.

Nous présentons dans la suite le spectrogramme d'un morceau de musique ainsi que celui de même morceau codé au format mp3 au débit de 64kbits/seconde.

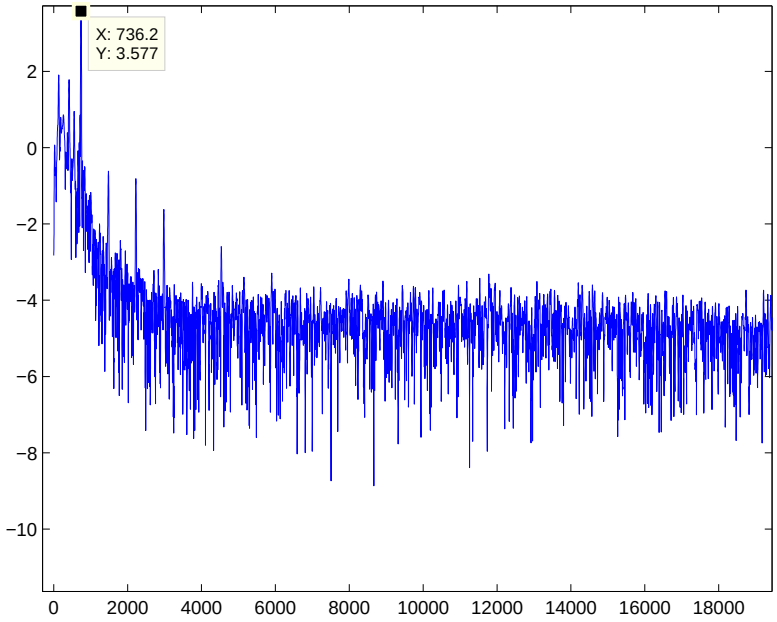


FIGURE 2.20 – **Une colonne du spectrogramme. C'est donc le contenu fréquentiel local autour d'un certain instant. Le pic le plus proéminent est pour la fréquence 736Hz, qui correspond à peu près à un Fa dièse.**

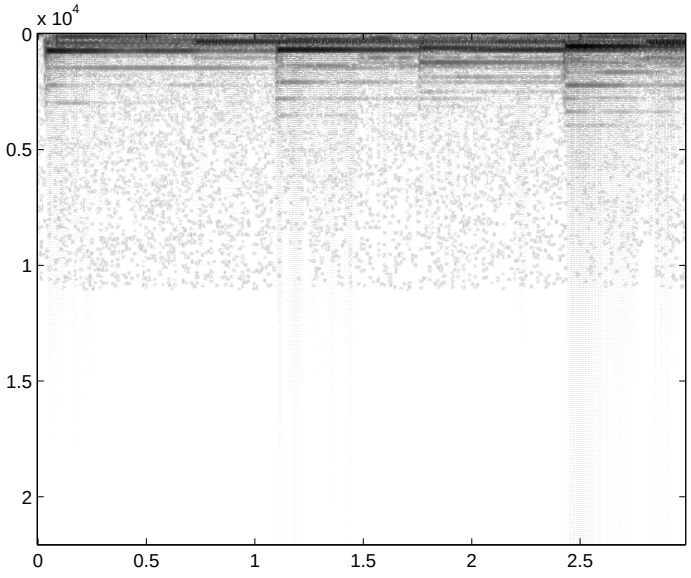


FIGURE 2.21 – **Spectrogramme du même morceau après qu'il ait été codé au format mp3. La remarque principale est la disparition de toutes les hautes fréquences au-delà de 10000Hz. Cependant le codage mp3 ne consiste pas seulement en l'application d'un filtre passe-bas. Il agit aussi par élimination de certaines fréquences qui sont cachées par d'autres. Il utilise pour cela des tables perceptuelles obtenues par des expériences de psychoacoustique qui indiquent les règles de masquage en fonction des fréquences et des amplitudes de deux ondes.**

Chapitre 3

Transformée en Z, les filtres discrets récursifs

Dans ce chapitre nous introduisons un nouvel outil, la transformée en Z, et étudions, grâce à ce nouvel outil, des propriétés fines des filtres discrets (SLI sur $\mathbb{Z}$).

3.1 Vocabulaire

Définition 3.1. *1. On dit qu'une suite h est* **causale** *si*

$$\forall n < 0,\ h_n = 0$$

2. On dit qu'un SLI est causal si sa réponse impulsionnelle est causale.

3. On dit qu'une suite h est **anti-causale** *si*

$$\forall n \geqslant 0,\ h_n = 0$$

4. On dit qu'un SLI est anti-causal si sa réponse impulsionnelle est anti-causale.

5. On dit qu'une suite est bilatère si elle n'est ni causale ni anticausale.

6. On dit qu'un SLI est à réponse impulsionnelle finie si sa réponse impulsionnelle est à support fini. On note en abrégé RIF.

7. On dit qu'un SLI est à réponse impulsionnelle infinie si sa réponse impulsionnelle n'est pas à support fini. On abrège en RII.

Proposition 3.2. *On a les propriétés suivantes qui sont toutes faciles à démontrer*

1. *La convolution de deux suites causale est causale.*
2. *La convolution d'une suite à support fini avec une autre suite à support fini est une suite à support fini*
3. *Et de même la composition de deux SLI causaux est causal et la composition de deux SLI RIF est RIF (car composer deux SLI revient à convoluer leurs réponses impulsionnelles).*

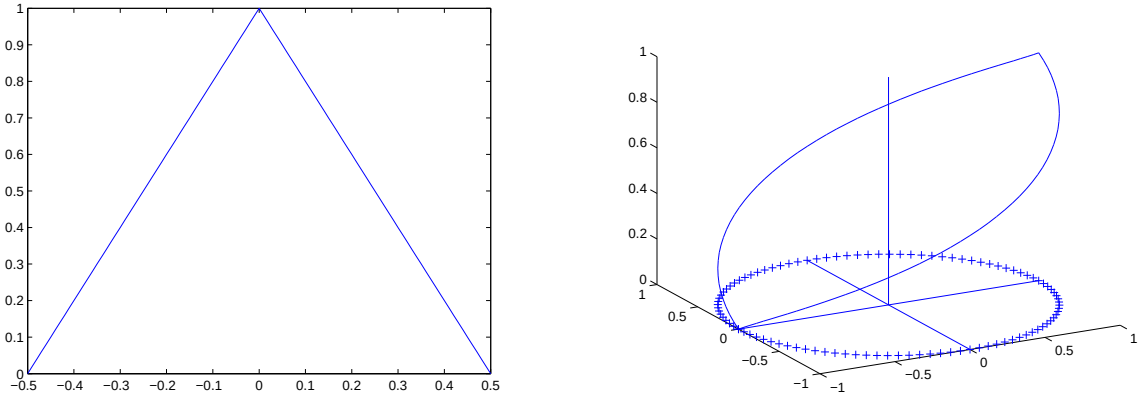


FIGURE 3.1 – TFtD d'une suite (à gauche) et sa TZ (à droite). Le cercle unité est figuré par les "+". L'axe vertical est celui de la valeur de la TZ. Le plan horizontal est le plan complexe. La courbe, continue da la TZ a un point haut en $z = 0$ qui correspond à la fréquence 0 et un point bas qui correspond au point $z = -1$ c'est-à-dire les fréquences $\frac{1}{2}$ et $-\frac{1}{2}$. On voit ici graphiquement le fait que les fréquences de $\mathbb{Z}$, $[-1/2, 1/2[$ doivent être vues comme périodiques.

3.2 Transformée en Z

Définition 3.3. *Si h est un signal défini sur $\mathbb{Z}$ et est sommable ($\sum |h_n| < +\infty$). On appelle transformée en Z de h, la fonction H définie sur le cercle unité de $\mathbb{C}$ (on note $\mathbb{U}$ le cercle unité $\mathbb{U} = \{z \in \mathbb{C} : |z| = 1\}$) par la formule :*

$$\forall z \in \mathbb{U}, H(z) = \sum_{n \in \mathbb{Z}} h_n z^{-n}.$$

La formule précédente a un sens, car la suite h_n est supposée sommable et que $|z| = 1 = |z^n|$.

On remarque que la transformée en Z s'obtient à partir de la transformée de Fourier à temps discret par un simple changement de variable. En effet, si $z \in \mathbb{U}$ alors $\exists \nu \in [-1/2, 1/2[$ tel que $z = e^{2i\pi\nu}$, si on note $\hat{h}$ la transformée de Fourier à temps discret de h alors on a

$$H(z) = H(e^{2i\pi\nu}) = \hat{h}(\nu). \tag{3.1}$$

La figure 3.1 montre à la fois la TFtD d'une suite et sa TZ.

Proposition 3.4. Théorème d'inversion appliqué à la transformée en Z et injectivité de la TZ
Si h est une suite sommable et que H est sa transformée en Z, alors on a

$$\forall n \in \mathbb{Z},\ h_n = \int_{-\frac{1}{2}}^{\frac{1}{2}} H(e^{2i\pi\nu}) e^{2i\pi n\nu} d\nu$$

Il s'agit simplement de l'application du théorème d'inversion combiné à l'équation 3.1.

En particulier si deux suites sommables ont la même transformée en Z, alors elles sont égales.

Exemple 3.5. **Transformée en Z des signaux finis**
Si h est une suite à support fini et causale, c'est-à-dire que les h_n sont nuls pour $n < 0$ et pour $n > N$ pour un certain N. Alors, la transformée en Z de h est un polynôme en z^{-1} dont la formule est

$$H(z) = \sum_{n=0}^{n \leqslant N} h_n z^{-n} = \sum_{n=0}^{n \leqslant N} h_n \left(z^{-1}\right)^n = P(z^{-1})$$

Où P est polynôme de degré N dont les coefficients sont les $h_n, n = 0 \dots N$.

Exemple 3.6. **Filtre exponentiel**
Soit h_n un signal défini par

$$h_n = \begin{cases} 0 \text{ si } n < 0 \\ \alpha^n \text{ sinon} \end{cases}$$

Avec $|\alpha| < 1$ (ce qui est nécessaire pour que h soit sommable). Alors, $H(z)$, la transformée en Z de h est donnée par la formule

$$\forall z \in \mathbb{U}, H(z) = \sum_{n \geqslant 0} (\alpha z^{-1})^n = \frac{1}{1 - \alpha z^{-1}}$$

La dernière égalité a lieu, car $|\alpha z^{-1}| = |\alpha| < 1$.

Ainsi, un signal d'extension infinie peut avoir une représentation très compacte grâce à la transformée en Z. C'est l'un des avantages de l'outil transformée en Z, il permet une analyse des filtres par l'analyse d'une fonction de la variable complexe dont l'expression peut être très simple.

Proposition 3.7. Lien entre transformée en Z et convolution

Si (x_n) et (y_n) sont deux signaux sommables et que leurs transformées en Z sont notées X et Y respectivement. On note u le signal $x * y$. On a déjà vu que u est sommable. Elle admet donc une transformée en Z notée U et celle-ci vérifie

$$\forall z \in \mathbb{U},\ U(z) = X(z)Y(z)$$

Démonstration.
Pour toute fréquence $\nu \in [-1/2, 1/2[$ on a (on note $\hat{x}$, $\hat{y}$ et $\hat{u}$ les transformées de Fourier à temps discret de x, y et u respectivement)

$$\hat{u}(\nu) = \hat{x}(\nu)\hat{y}(\nu)$$

Or d'après l'équivalence vue à l'équation 3.1, si $z = e^{2i\pi\nu}$ l'équation précédente devient

$$U(z) = \hat{u}(\nu) = \hat{x}(\nu)\hat{y}(\nu) = X(z)Y(z)$$

□

3.3 Les filtres (SLI) récursifs

Dans cette partie nous étudions les filtres récursifs, on utilise la transformée en Z pour en étudier les propriétés.

Définition 3.8. Filtres récursifs stables
Un SLI sur $\mathbb{Z}$ est dit récursif stable si il vérifie les conditions suivantes

1. *Sa réponse impulsionnelle est sommable ($\sum_n |h_n| < +\infty$).*
2. *Il existe des coefficients $a_0 \dots a_p$ et $b_0 \dots b_q$ tels que si (x_n) est une entrée et (y_n) la sortie qui lui correspond par le SLI en question alors*

$$\forall n \in \mathbb{Z}, b_0 y_n + b_1 y_{n-1} + \cdots + b_q y_{n-q} = a_0 x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p}$$

   Les a_i et b_j sont appelés coefficients du SLI.
3. *On suppose de plus que les polynômes*

$$\sum b_i z^i$$

   et

$$\sum a_i z^i$$

   sont premiers entre eux. C'est-à-dire qu'ils n'ont aucun zéro commun (en tant que polynômes sur $\mathbb{C}$). Cette condition technique est là pour éviter des cas pathologiques qui n'ont pas d'intérêt.

Remarque 3.9. On appelle ces filtres "récursifs", car la manière de les implémenter est récursive. On verra dans la suite à quelles conditions l'implémentation récursive donne la bonne solution.

Les deux premières conditions ne sont pas *a priori* compatibles entre elles. Dans ce qui suit, nous montrerons la condition nécessaire et suffisante sur les coefficients a_i et b_j pour que les deux conditions soient compatibles. D'abord, nous donnons l'expression de la transformée en Z de la réponse impulsionnelle d'un tel filtre (ce qui est équivalent à donner sa TFtd).

Proposition 3.10. Transformée en Z de la réponse impulsionnelle d'un filtre récursif stable
Si T est un SLI récursif stable dont les coefficients sont les $a_0 \dots a_p$ et $b_0 \dots b_q$ et que l'on note h sa réponse impulsionnelle. Alors, h admet une transformée en Z (car sommable par hypothèse) et sa transformée en Z, notée H est donnée par

$$H(z) = \frac{P(z^{-1})}{Q(z^{-1})}$$

où P et Q sont des polynômes dont les coefficients sont les a et les b respectivement

$$P(t) = \sum_{i=0}^{i=p} a_i t^i$$

$$Q(t) = \sum_{i=0}^{i=q} b_i t^i$$

En particulier Q n'a pas de zéros sur $\mathbb{U}$.

Démonstration.
On s'intéresse au cas où x est l'impulsion en 0 ($x_n = \delta_n^0$). On sait alors que y (la sortie correspondante) est la réponse impulsionnelle h et on a, par hypothèse sur le filtre

$$\forall n \in \mathbb{Z}, b_0 h_n + b_1 h_{n-1} + \cdots + b_q h_{n-q} = a_0 \delta_n^0 + a_1 \delta_{n-1}^0 + \cdots + a_p \delta_{n-p}^0$$

Si on note g_n la suite de gauche et d_n la suite de droite de l'égalité ci-dessus, c'est-à-dire :

$$g_n = b_0 h_n + b_1 h_{n-1} + \cdots + b_q h_{n-q}$$

et

$$d_n = a_0 \delta_n^0 + a_1 \delta_{n-1}^0 + \cdots + a_p \delta_{n-p}^0$$

La suite g est la convolution de la réponse impulsionnelle h avec le signal à support fini $b_0, \ldots, b_q$. D'après la proposition 3.7 la suite g possède bien une transformée en Z et celle-ci vaut (on la note G)

$$\forall z \in \mathbb{U},\ G(z) = Q(z^{-1})H(z)$$

(Le polynôme Q est défini dans le corps de la proposition). De même la séquence d est la convolution de l'impulsion δ^0 avec le signal à support fini $a_0, \ldots, a_p$. Comme la convolution de l'impulsion avec un autre signal laisse ce signal inchangé, la suite d_n est simplement a_n. Elle a donc pour transformée en Z,

$$\forall z \in \mathbb{U},\ D(z) = P(z^{-1})$$

Comme les suite g et d sont égales, elles ont la même transformée en Z. On a donc

$$\forall z \in \mathbb{U},\ H(z)Q(z^{-1}) = P(z^{-1})$$

H est bien définie sur le cercle unité, car h est sommable. Le polynôme Q ne peut avoir de zéro sur le cercle unité, car sinon P devrait avoir le même zéro (en raison de l'équation ci-dessus). Or, cela est exclu par notre définition des filtres récursifs stables. On a donc,

$$\forall z \in \mathbb{U},\ H(z) = \frac{P(z^{-1})}{Q(z^{-1})}$$

□

Proposition 3.11. Décomposition d'une fraction rationnelle sur $\mathbb{C}$
Soit P et Q deux polynômes sur $\mathbb{C}$ premiers entre eux et tels que Q n'a pas de zéros sur $\mathbb{U}$. Alors il existe deux réels R_1 et R_2, des constantes réelles C_1 et C_2, et une suite h tels que

$$\begin{array}{l} 0 < R_1 < 1 < R_2 \\ \forall n \geqslant 0,\ |h_n| < C_1 R_1^n \text{ et } |h_{-n}| < C_2 R_2^{-n} \\ \forall R_1 < |z| < R_2,\ \frac{P(z^{-1})}{Q(z^{-1})} = \sum_{n\in\mathbb{Z}} h_n z^{-n} \end{array}$$

Ainsi h est sommable et

$$\forall z \in \mathbb{U},\ H(z) = \frac{P(z^{-1})}{Q(z^{-1})}$$

Cette proposition n'est pas démontrée.

Exemple 3.12. **Important** Les décompositions qui suivent sont très importantes pour comprendre le rôle que jouent les pôles dans la causalité d'un SLI récursif.

On a, si $|\alpha| < 1$,

$$\forall z \in \mathbb{U},\ \frac{1}{1 - z^{-1}\alpha} = \sum_{n \geqslant 0} \alpha^n z^{-n}$$

Ce qui signifie que la fonction ci-dessus est la TZ de la suite causale de terme α^n pour n positif ou nul. et si $|\alpha| > 1$

$$\forall z \in \mathbb{U},\ \frac{1}{1 - z^{-1}\alpha} = -\alpha^{-1} z \frac{1}{1 - \alpha^{-1} z} = -\sum_{n>0} \alpha^{-n} z^n = -\sum_{m<0} \alpha^m z^{-m}$$

Ce qui signifie que la fonction ci-dessus est la TZ de la suite anticausale de terme général $-\alpha^m$ pour m strictement négatif.

Proposition 3.13. Unicité de la solution d'une équation de récurrence
On se donne une équation de récurrence

$$\forall n \in \mathbb{Z}, b_0 y_n + b_1 y_{n-1} + \cdots + b_q y_{n-q} = a_0 x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p}$$

Telle que le polynôme

$$Q(t) = \sum_{i=0}^{i=q} b_i t^i$$

n'a pas de zéros sur le cercle unité, $\mathbb{U}$*. Et on suppose aussi, que le polynôme* $P(t) \sum_{i=0}^{i=p} a_i t^i$ *est premier avec* Q*.*

Alors pour toute suite **sommable** x *il existe une unique suite* **sommable** y *qui vérifie l'équation de récurrence. Cette solution est de la forme*

$$y = h * x$$

où h *est l'unique solution sommable de l'équation de récurrence lorsque* x *est l'impulsion en zéro. La TZ de* h *a pour équation*

$$\forall z \in \mathbb{U},\ H(z) = \frac{P(z^{-1})}{Q(z^{-1})}$$

Ainsi, une équation de récurrence définit un SLI des suites sommables vers les suites sommables (avec la condition sur les zéros de Q*).*

Remarque 3.14. Sans l'hypothèse y sommable, nous pouvons trouver une infinité de solutions pour chaque entrée x et ce dès que le polynôme Q n'est pas une constante.

Démonstration.
Unictié
Soit x une suite sommable et y^1 et y^2 deux solutions (sommables) de l'équation de récurrence (la notation en exposant y^1 et y^2 ne signifie pas, ici, qu'elles sont les translatées d'une même suite). En éliminant le terme de droite de l'équation de récurrence on a

$$\forall n \in \mathbb{Z},\ b_0 y^1_n + b_1 y^1_{n-1} + \cdots + b_q y^1_{n-q} = b_0 y^2_n + b_1 y^2_{n-1} + \cdots + b_q y^2_{n-q}$$

Si on note Y^1 et Y^2 les transformées en Z de y^1 et y^2, l'équation ci-dessus se réécrit

$$\forall z \in \mathbb{U},\ Q(z^{-1})Y^1(z) = Q(z^{-1})Y^2(z)$$

Et comme Q n'a pas de zéros sur le cercle unité

$$\forall z \in \mathbb{U},\ Y^1(z) = Y^2(z)$$

D'où, $y^1 = y^2$ car la transformée en Z est injective (comme la TFtD).
Forme de la solution : $y = x * h$

On appelle h la suite sommable telle que sa TZ, notée H vérifie :

$$\forall z \in \mathbb{U},\ H(z) = \frac{P(z^{-1})}{Q(z^{-1})}$$

D'après 3.11 une telle suite existe et est sommable.

Soit x une suite sommable, X sa TZ. Soit $y = h * x$. y est sommable (règles de calcul) et sa TZ est

$$\forall z \in \mathbb{U}\ Y(z) = H(z)X(z)$$

et donc

$$\forall z \in \mathbb{U}\ Q(z^-1)Y(z) = P(z^{-1})X(z)$$

Or, la convolution de la suite finie $b_0, \ldots, b_q$ avec y a pour TZ

$$Q(z^-1)Y(z)$$

De même, la convolution de la suite finie $a_0, \ldots a_p$ avec y a pour TZ

$$P(z^{-1})X(z)$$

Par injectivité de la TZ, on en déduit que les deux suites sont égales. Or l'égalité de ces deux suites est exactement l'expression de l'équation de récurrence. Par ailleurs si on prend pour x la suite impulsion, on voit que h est bien la sortie qui lui est associée (pour la preuve nous avons défini h par sa TZ). □

Simplification d'une équation de récurrence
Si x et y vérifient l'équation

$$\forall n \in \mathbb{Z}, b_0y_n + b_1y_{n-1} + \cdots + b_qy_{n-q} = a_0x_n + a_1x_{n-1} + \cdots + a_px_{n-p}$$

On ne s'intéresse jamais au cas où tous les b_j sont nuls, car sinon le polynôme Q est nul et a donc des zéros sur $\mathbb{U}$. Si tous les a_i sont nuls, la seule solution à l'équation de récurrence est $y = 0$, là encore ce cas est peu intéressant.

Si $b_0 = 0$ il suffit de remplacer la suite y par la suite $y^1_n = y_{n-1}$ (y^1 est la 1-translatée de y) pour avoir

$$\forall n \in \mathbb{Z}, b_1y^1_n + b_2y^1_{n-1} + \cdots + b_qy^1_{n-q+1} = a_0x_n + a_1x_{n-1} + \cdots + a_px_{n-p}$$

On peut faire de même pour la suite x si $a_0 = 0$.

Ainsi, on peut, quitte à translater les suites x et y, supposer que les équations de récurrence ont leurs coefficients a_0 et b_0 non nuls. Par ailleurs, en divisant les deux parties de l'équation par b_0 on peut aussi supposer que $b_0 = 1$. Enfin, quitte à remplacer la suite x par la suite a_0x et diviser tous les a_i par a_0, on peut aussi supposer que $a_0 = 1$.

Définition 3.15. Pôles et zéros
Soit un filtre récursif stable dont l'équation de récurrence est

$$\forall n \in \mathbb{Z}, b_0 y_n + b_1 y_{n-1} + \cdots + b_q y_{n-q} = a_0 x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p}$$

et P et Q les polynômes

$$P(t) = \sum_{i=0}^{i=p} a_i t^i$$

$$Q(t) = \sum_{i=0}^{i=q} b_i t^i$$

On suppose de plus que a_0 et b_0 sont non nuls.

On appelle **zéros** *du filtre les zéros de la fonction $z \mapsto P(z^{-1})$ c'est-à-dire les inverses des zéros du polynôme P (remarquer que $P(0) = a_0 \neq 0$). Ils sont en nombre p (comptés avec multiplicité).*

On appelle **pôles** *du filtre les zéros e la fonction $z \mapsto Q(z^{-1})$ c'est-à-dire les inverses des zéros du polynôme Q (remarquer que $Q(0) = b_0 \neq 0$). Ils sont en nombre q (comptés avec multiplicité).*

Proposition 3.16. Inversion des filtres récursifs stables
Si un filtre récursif stable T_1 dont la relation de récurrence est

$$\forall n \in \mathbb{Z}, b_0 y_n + b_1 y_{n-1} + \cdots + b_q y_{n-q} = a_0 x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p}$$

n'a pas de zéros sur le cercle unité (P ne s'annule pas sur $\mathbb{U}$) alors son filtre inverse est récursif stable T_2 dont l'équation de récurrence est

$$\forall n \in \mathbb{Z}, a_0 y_n + a_1 y_{n-1} + \cdots + a_p y_{n-p} = b_0 x_n + b_1 x_{n-1} + \cdots + b_q x_{n-q}$$

Par ailleurs, les pôles de T_2 sont les zéros de T_1 et les zéros de T_2 sont les pôles de T_1.

Démonstration.
Si H_1 est la TZ de la réponse impulsionnelle du filtre T_1 et H_2 la TZ de la réponse impulsionnelle de T_2 (tel que défini dans le corps de la proposition). T_2 est bien défini, car P n'a pas de zéros sur le cercle unité. On a, par ce qui précède :

$$H_1(z) = \frac{P(z^{-1})}{Q(z^{-1})}$$

et

$$H_2(z) = \frac{Q(z^{-1})}{P(z^{-1})} = \frac{1}{H_1(z)}$$

Si x est une suite sommable et $y = T_1(x)$ la sortie associée. La suite y est bien sommable (comme convolution de x avec la réponse impulsionnelle de T_1 qui est sommable). On note $w = T_2(y)$ la sortie associée par T_2 à la suite y.

La TZ de w, notée W, est

$$W(z) = H_2(z)Y(z) = H_2(z)\left(H_1(z)X(z)\right) = X(z)$$

Donc, w et x ont la même TZ, sont égales par injectivité de la TZ et T_2 est bien le filtre inverse de T_1. □

Proposition 3.17. Module de la TZ en fonction des pôles et des zéros du filtre *Soit un filtre récursif dont la TZ de la réponse impulsionnelle est*

$$H(z) = \frac{P(z^{-1})}{Q(z^{-1})}$$

Avec

$$P(t) = \sum_{i=0}^{i=p} a_i t^i \text{ et } Q(t) = \sum_{i=0}^{i=q} b_i t^i$$

On fait l'hypothèse que $a_0 = b_0 = 1$ *et* $a_p \neq 0$ *et* $b_q \neq 0$ [1]. *On note* $\alpha_i^{-1},\ i = 1, \ldots, p$ *les zéros de* P *et* $\beta_j^{-1},\ j = 1, \ldots, q$ *les zéros de* Q. *Ces zéros sont comptés avec multiplicité et on a pu les regarder comme les inverses des* α_i *et,* β_i *car* $P(0) = a_0 \neq 0$ *et* $Q(0) = b_0 \neq 0$.

Alors on a la propriété suivante pour la norme de $H(z)$ *:*

$$\forall z \in \mathbb{U},\ |H(z)| = \frac{\prod_{i=1}^{p} |1 - z^{-1}\alpha_i|}{\prod_{j=1}^{q} |1 - z^{-1}\beta_i|} = \frac{\prod_{i=1}^{p} |z - \alpha_i|}{\prod_{j=1}^{q} |z - \beta_i|}$$

Et les α_i *sont les zéros de ce filtre et les* β_i *sont les pôles du filtre.*

Démonstration.
On a $P(z) = a_p \prod_i (z - \alpha_i^{-1})$, en particulier $1 = a_0 = P(0) = a_p \prod_i (-\alpha_i)^{-1}$. En remplaçant par z^{-1}, on a

$$P(z^{-1}) = a_p \prod_i (z^{-1} - \alpha_i^{-1}) = a_p \prod_i (-\alpha_i)^{-1} \prod_i (1 - z^{-1}\alpha_i) = \prod_i (1 - z^{-1}\alpha_i)$$

$$|P(z^{-1})| = \prod_i |1 - z^{-1}\alpha_i|$$

De plus sur $\mathbb{U}$ on a

$$\forall z \in \mathbb{U},\ |P(z^{-1})| = \prod_i |1 - z^{-1}\alpha_i| = \prod_i |z||1 - z^{-1}\alpha_i| = \prod_i |z - \alpha_i|$$

(car $|z| = 1$ sur $\mathbb{U}$)

On fait de même pour Q

$$\forall z \in \mathbb{U},\ |Q(z^{-1})| = \prod_j |1 - z^{-1}\beta_j| = \prod_j |z - \beta_i|$$

Et comme $|H(z)|$ est le rapport des modules $P(z^{-1})$ et $Q(z^{-1})$, on obtient le résultat annoncé.

□

On peut ainsi interpréter le module de $H(z)$ comme suit : Si M est un point sur le cercle unité dont l'affixe est z. On note A_i les points dont les affixes sont les α_i (les zéros

1. On a vu plus haut comment faire pour se ramener depuis une équation de récurrence quelconque à ces conditions.

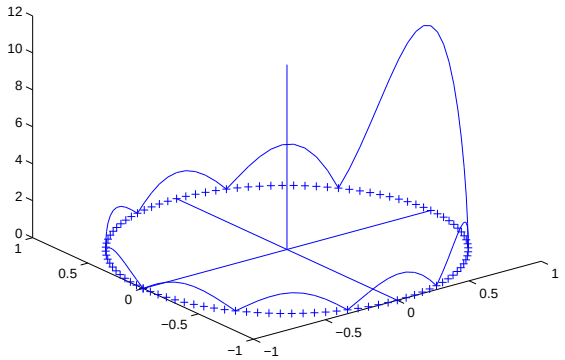


FIGURE 3.2 – Module de la TZ de la suite constante égale à 1 entre 0 et $N-1$ et nulle ailleurs (RI de la moyenne glissante). Ici $N = 10$. On voit bien que la TZ passe par zéro à toutes les racines N-ièmes de l'unité sauf la racine 1.

d'un filtre récursif) et B_j les points dont les affixes sont les β_j les pôles de ce même filtre. Alors le module de $H(z)$, la TZ de la réponse impulsionnelle de ce filtre, est donné par

$$|H(z)| = \frac{\prod_i MA_i}{\prod_j MB_j}$$

(si C et D sont deux points du plan, CD signifie la longueur du segment qui joint C et D.)

Ainsi, le module de $H(z)$ est vu comme le rapport entre le produit des distances aux zéros et le produit des distances aux pôles. Plus z s'approche d'un pôle, plus le module de $H(z)$ est grand. Plus z s'approche d'un zéro, plus le module de $H(z)$ est petit (et même nul si z est un zéro du filtre).

Exemple 3.18. Si $h_n = 1$ pour $n = 0, \ldots, N-1$ et $h_n = 0$ pour les autres valeurs de n. On a

$$\forall z \in \mathbb{U},\ H(z) = \sum_{k=0}^{N-1} z^{-k} = \prod_{k=1}^{N-1} (1 - z^{-1}\omega_N^k) \text{ avec } \omega_N = e^{2i\pi\frac{1}{N}}$$

(le polynôme $1+z+z^2+\cdots+z^{N-1}$ a pour racines toutes les racines N-ièmes de l'unité sauf la racine 1).

On comprend par l'interprétation ci-dessus, pourquoi la TFtD de h a la forme particulière qu'on lui connaît. Lorsque ν parcourt $[-1/2, 1/2[$, $\hat{h}(\nu) = H(e^{2i\pi\nu})$ s'annule à chaque fois que que $z = e^{2i\pi\nu}$ est une racine N-ième de l'unité (sauf pour $\nu = 0$ qui correspond à $z = 1$). La figure 3.2 montre bien ce phénomène.

Proposition 3.19. Causalité d'un filtre récursif
Un SLI récursif stable dont l'équation de récurrence est

$$\forall n \in \mathbb{Z}, y_n + b_1 y_{n-1} + \cdots + b_q y_{n-q} = a_0 x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p}$$

(notez qu'on a choisi $b_0 = 1$ ici) est causal si et seulement si tous ses pôles sont strictement dans le disque unité (i.e. de module strictement plus petit que 1)

Démonstration.
Nous faisons la preuve seulement dans le cas où les pôles du filtre ont une multiplicité 1. La preuve dans le cas général est seulement un peu plus fastidieuse et calculatoire. On suppose donc que Q n'a que des racines simples.

Si on note β_j, $j = 1, \ldots, q$ les pôles du SLI récursif alors il existe un polynôme R est des constantes complexes γ_j (dont aucune n'est nulle) tels que la réponse impulsionnelle du SLI récursif a pour TZ

$$H(z) = \frac{P(z^{-1})}{Q(z^{-1})} = R(z^{-1}) + \sum_j \frac{\gamma_j}{1 - z^{-1}\beta_j}$$

Cette écriture résulte de la décomposition en éléments simples d'une fraction rationnelle. Si l'un des γ_j était nul, alors le pôle correspondant ne serait pas un zéro de $Q(z^{-1})$.

Comme R est un polynôme et que l'on note $r_0, \ldots, r_k$ ses coefficients, il est équivalent que h soit causale ou que $h - r$ soit causale. La TZ de $h - r$ que l'on note H_1 est

$$H_1(z) = \sum_j \frac{\gamma_j}{1 - z^{-1}\beta_j}$$

1. Si on suppose que tous les β_j sont de module < 1 alors, d'après l'exemple 3.12 H_1 est bien la TZ d'une suite causale.
2. Si on suppose que $|\beta_1| > 1$ et $|\beta_2| > 1 \ldots |\beta_l| > 1$ et que les β_m pour $m > l$ vérifient $|\beta_m| < 1$ (c'est-à-dire que l'on suppose que les $l > 0$ premiers β sont de module strictement plus grand que 1). D'après l'exemple 3.12 pour $n < 0$ la suite h_n vaut

   $$h_n = -\sum_{j=1}^{l} \gamma_j (\beta_j)^n$$

   pour n assez grand (en valeur absolue) le β de module le plus petit domine les autres et h_n ne peut être nul. (en élevant à une puissance assez grande la différence des modules compense les facteurs γ_j). Donc h ne peut être causale.

On a donc prouvé l'équivalence entre le la caractère causal de h et le fait que tous les pôles du filtre soient de module strictement inférieur à 1.

La preuve dans le cas général (c'est-à-dire avec des pôles de multiplicité plus grande que 1) aurait été simplement plus fastidieuse, il aurait fallu considérer la décomposition de facteurs du type

$$\frac{1}{(1 - z^{-1}\beta)^k}$$

qui ont pour décomposition sur la base des z^{-n} des facteurs du type $G(n)\alpha^n$ où G est un polynôme de degré $k - 1$.

□

Définition 3.20. Filtres à minimum de phase
Un filtre récursif stable est dit à minimum de phase s'il est causal et que son inverse est aussi stable et causal. Par ce qui précède cela est équivalent à dire que ses pôles et ses zéros sont à l'intérieur du disque unité (car les zéros d'un filtre sont les pôles de son inverse).

3.3.1 Implémentation des filtres récursifs

On se demande comment implémenter un filtre récursif stable. La proposition suivante montre que si le filtre est causal, une implémentation intuitive permet de résoudre l'équation de récurrence et ce de manière exacte si l'entrée est aussi causale ou approchée si la suite en entrée n'est pas causale.

Dans le cas où le filtre n'est pas causal, nous donnons des exemples où l'implémentation proposée renvoie une réponse non seulement fausse mais aussi numériquement instable.

Proposition 3.21. *Si T est un SLI récursif stable dont l'équation de récurrence est de la forme*

$$\forall n \in \mathbb{Z}, y_n + b_1 y_{n-1} + \cdots + b_q y_{n-q} = a_0 x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p}$$

Et que de plus T est supposé **causal** *(voir la proposition 3.19).*

Si x est une suite sommable et $y = T(x)$ la seule solution sommable de l'équation de récurrence. On note x^c la suite définie par

$$x_n^c = \begin{cases} 0 \text{ si } n < 0 \\ x_n^c = x_n \text{ si } n \geqslant 0 \end{cases}$$

(x^c peut être appelée troncature causale de x)

On considère la suite causale t définie par

$$t_n = \begin{cases} 0 \text{ si } n < 0 \\ a_0 x_n^c + a_1 x_{n-1}^c + \cdots + a_p x_{n-p}^c - (b_1 t_{n-1} + \cdots + b_q t_{n-q}) \text{ si } n \geqslant 0 \end{cases}$$

alors on a

1. *Si x est causale alors $t = y$. Autrement dit cette implémentation est parfaite.*
2. *Dans tous les cas*

$$\exists A < 1,\ C \geqslant 0,\ \forall n \geqslant 0,\ |t_n - y_n| < CA^n \|x\|_1$$

Autrement dit, pour n assez grand, t devient aussi proche que l'on veut de la vraie solution y.

Remarque 3.22.
Cette proposition simule bien ce que fait un filtre réaliste, qui commence à analyser un signal à partir d'un instant donné. Elle dit que si le signal commence au même moment, alors l'implémentation est parfaite et que sinon, notre implémentation approche d'aussi près que voulu la solution théorique (qui aurait nécessité l'observation de tout le passé de x).

Démonstration. On appelle h la réponse impulsionnelle de du SLI T. On sait que h et sommable et $y = h * x$. Par ailleurs on a supposé que h est causale et la proposition 3.11 nous dit que

$$\exists R < 1,\ C,\ \forall n \geqslant 0,\ |h_n| \leqslant CR^n$$

1. Si x est causale :
   La solution $y = h * x$ est aussi causale (la convolution de deux suites causales est causale). Donc on a
   $$\forall n < 0,\ y_n = 0 = t_n$$
   On va montrer par récurrence sur $n \geqslant -1$ que $y_n = t_n$. On vient de montrer cela pour tous les $n < 0$ en particulier pour $n = -1$. Il nous reste à faire le passage de $n-1$ à n avec $n \geqslant 0$ Comme y est solution de l'équation de récurrence on a
   $$y_n = a_0 x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p} - (b_1 y_{n-1} + \cdots + b_q y_{n-q}) \text{ si } n \geqslant 0$$
   Par hypothèse de récurrence on a
   $$\forall j = 1, \ldots, q,\ y_{n-j} = t_{n-j}$$
   Donc
   $$\begin{aligned} y_n &= a_0 x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p} - (b_1 y_{n-1} + \cdots + b_q y_{n-q}) \\ &= a_0 x^c_n + a_1 x^c_{n-1} + \cdots + a_p x^c_{n-p} - (b_1 t_{n-1} + \cdots + b_q t_{n-q}) = t_n \end{aligned}$$
   Comme x est causale $x = x^c$ et la dernière égalité est due à la définition de t_n pour n positif ou nul. On a donc bien montré $t = y$.
2. On a encore $y = h * x$. Par ailleurs, il est facile de voir que
   $$t = x^c * h$$
   En effet, t ne dépend que des x_n pour n positif. En remplaçant x par x^c on obtient donc la même suite t. Mais x^c est causale, donc, par ce qui précède, t est exactement la sortie associée à x^c qui est $x^c * h$. Donc $y - t = (x - x^c) * h$ et
   $$\forall n \in \mathbb{Z},\ y_n - t_n = \sum_{m \in \mathbb{Z}} (x_m - x^c_m) h_{n-m} = \sum_{m<0} (x_m - x^c_m) h_{n-m} = \sum_{m<0} x_m h_{n-m}$$
   Car $x_m - x^c_m$ est nul pour $m \geqslant 0$ et x^c_m est nul pour $m < 0$. Pour n positif on a
   $$|y_n - t_n| \leqslant \sum_{m<0} |x_m||h_{n-m}| \leqslant \sum_{m<0} |x_m| C R^n = \|x\|_1 C R^n$$
   La seconde inégalité est due au fait que $m < 0 \Longrightarrow n - m > n$ et donc $R^{n-m} \leqslant R^n$. On a donc prouvé le résultat avec $A = R$.

□

Exemple 3.23. **Implémentation rapide de l'inverse d'un filtre RIF**
On se donne un SLI dont la réponse impulsionnelle est $h_0 = 1,\ h_1 = \frac{1}{2}$ (et h_n pour les autres valeurs de n). Si x est une suite et y la sortie qui lui est associée par ce SLI, on a
$$\forall n \in \mathbb{Z} y_n = (h * x)_n = x_n + 1/2 x_{n-1}$$
Ce SLI est donc stable et récursif (comme tout SLI qui est à RIF). Il est aussi causal. Son SLI inverse est donc récursif avec comme équation de récurrence (si x est l'entrée et y la sortie)
$$y_n + 1/2 y_{n-1} = x_n$$

Le seul pôle de ce filtre est $-\frac{1}{2}$ (seul zéro de la fonction $z \mapsto 1 + \frac{1}{2}z^{-1}$). Comme $-1/2$ est de module plus petit que 1. Ce SLI peut être implémenté de manière récursive causale comme ci-dessus.

De plus la réponse impulsionnelle, notée g, de ce filtre est $g_n = \left(-\frac{1}{2}\right)^n$ pour $n \geqslant 0$.

Si nous devions l'implémenter en appliquant la formule de convolution, il nous faudrait une infinité d'opérations par échantillon (car g est à support infini), alors qu'il suffit de deux opérations par échantillon pour l'implémenter suivant la procédure récursive.

Exemple 3.24. **Instabilité lors de l'implémentation d'un filtre récursif non causal** Considérons le filtre récursif dont l'équation de récurrence est

$$\forall n \in \mathbb{Z},\ y_n - 2y_{n-1} = x_n$$

Sa réponse impulsionnelle est $n \mapsto -(2)^n$ pour $n < 0$ et zéro sinon.

Essayons de lui appliquer l'algorithme de la proposition 3.21 avec comme entrée l'impulsion $x_n = \delta_n^0$ (donc $x^c = x$). On pose donc $t_n = 0$ pour $n < 0$ et $t_n = x_n + 2t_{n-1}$ pour $n \geqslant 0$. Il est facile de voir que

$$t_n = \begin{cases} 0 \text{ si } n < 0 \\ 2^n \text{ si } n \geqslant 0 \end{cases}$$

Ainsi, l'application de la méthode récursive pour appliquer un filtre qui ne serait pas causal, ne peut pas fonctionner.

On pourrait remarquer qu'en modifiant t_{-1} par $t_{-1} = -\frac{1}{2}$ on aurait comme résultat de l'implémentation $t_n = 0$ pour tout $n \geqslant 0$. Mais ceci n'est pas une solution pour le cas général où x serait plus compliqué qu'une impulsion (si on voulait généraliser, cela reviendrait à attendre la fin du signal x et appliquer une convolution avec la réponse impulsionnelle anticausale). Par ailleurs, même dans ce cas simple (et en remplaçant 2 par $\sqrt{\pi}$ pour que la représentation informatique ne soit pas parfaite) on constate quand même une explosion numérique de ce procédé. Voici un exemple en matlab qui montre ce phénomène.

```
>> deux=sqrt(pi); %on remplace 2 par racine(pi) pour montrer l'instabilité
>> P=[1]; %le polynome constat egale à 1
>> Q=[1 -deux]; % le polynome 1-2z
>> t=filter(P,Q,[-1/deux 1 zeros(1,100)]); %applique un filtre récursif de TZ P/Q
>> t(6)

   1.0957e-15

>> t(101)

   4.5116e+08
```

On applique notre algorithme récursif par la commande matlab "filter" qui fait ce qui est décrit dans la proposition 3.21, en utilisant l'astuce de la modification de t_{-1}. Une première fois on laisse l'algorithme se dérouler 6 fois (on calcule t_5) et on a t_5 de l'ordre de 10^{-15} ce qui est compatible avec notre prédiction ($t_n = 0$ pour n positif). Mais si on laisse le calcul se dérouler plus de 100 fois alors le dernier t_n est de l'ordre de 10^8 au lieu de zéro !

3.3.2 Introduction à la synthèse de filtre

Nous allons illustrer sur un exemple, les méthodes de synthèse de filtre. Une synthèse de filtre a pour but de construire un filtre pour réaliser une tâche précise. Ici, nous choisissons le filtre passe-bas parfait dont la RI est notée h et la réponse fréquentielle, $\hat{h}$, est donnée par :

$$\forall \nu \in [-1/2, 1/2[,\ \hat{h}(\nu) = \begin{cases} 0 \text{ si } |\nu| > \frac{1}{8} \\ 1 \text{ si } |\nu| \leqslant \frac{1}{8} \end{cases}$$

$\hat{h}$ est d'énergie finie et le théorème d'inversion nous dit que h ne peut-être que :

$$h_n = \int_{-\frac{1}{2}}^{\frac{1}{2}} \hat{h}(\nu) e^{2i\pi\nu n} d\nu = \int_{-\frac{1}{8}}^{\frac{1}{8}} e^{2i\pi\nu n} d\nu = \frac{\sin\left(\frac{\pi}{4}n\right)}{\pi n} \text{ (et } 1/4 \text{ si } n = 0)$$

Méthode de la fenêtre Dans la méthode de la fenêtre, on se limite à approximer h par une suite à support fini dont on fixe par avance le nombre d'échantillons non nuls (et ce pour des raisons de possibilité d'implémentation pratique). Par exemple, on va forcer g à vérifier ceci, pour N fixé :

$$g_n = 0 \text{ si } |n| > N$$

Ainsi, g a au plus $2N + 1$ valeurs non nulles.

On prend pour critère de bonne approximation, mais d'autres choix sont possibles, que la TFtD de g approche au mieux la TFtD de h au sens des moindres carrés :

$$\text{on veut minimiser } \int_{-\frac{1}{2}}^{\frac{1}{2}} |\hat{h}(\nu) - \hat{g}(\nu)|^2 d\nu$$

Par l'égalité de Parseval ($\hat{h} - \hat{g}$ est la TFtD de $h - g$) cela revient à minimiser

$$\sum_n |h_n - g_n|^2 = \sum_{|n|>N} |h_n|^2 + \sum_{|n|\leqslant N} |g_n - h_n|^2$$

Le premier terme de la somme ne peut être modifié par le choix des valeurs des g_n. Le second peut être rendu nul en choisissant $g_n = h_n$.On aboutit donc au choix

$$g_n = \begin{cases} 0 \text{ si } n > N \\ h_n = \frac{\sin\left(\frac{\pi}{4}n\right)}{\pi n} \text{ si } n \leqslant N \end{cases}$$

SI on choisi N=7 (soit 15 coefficients), on a une erreur quadratique de $\sum_n |h_n - g_n|^2 = 0.0125$ et plus généralement cette erreur quadratique décroît comme $1/N$.

Synthèse par choix des pôles et des zéros

D'après l'interprétation du module de la TZ comme le rapport entre distance aux zéros et distances aux pôles, une autre possibilité pour la synthèse de filtre est de choisir des pôles près de la zone où l'on veut que la TZ soit élevée et les zéros près de la zone où l'on veut que la TZ soit petite. Si nous voulons que l'implémentation causale soit stable il faut choisir tous les pôles à l'intérieur du disque unité. Il y a différentes méthodes pour choisir les pôles et les zéros, mais ces algorithmes dépassent le cadre de ce cours.

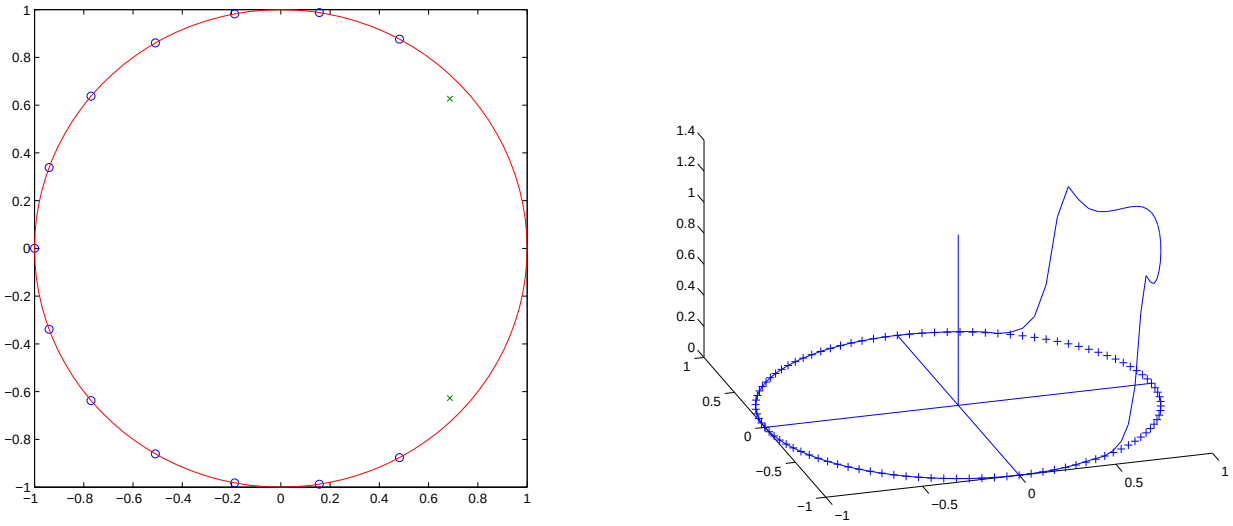


FIGURE 3.3 – À gauche position des pôles (figurés par des 'x') et des zéros (figurés par des 'o') et ce par rapport au cercle unité. À droite : Avec ce choix de pôles et de zéros, nous obtenons le module de la TZ figuré ici.

Disons simplement, que pour se comparer à la méthode de la fenêtre, on va se fixer un nombre d'opérations qui est le total du nombre de pôles et de zéros plus 1. En effet, pour chaque calcul d'un terme de t_n dans la proposition 3.21 on effectue $p+q-1$ multiplications et $p+q-2$ additions.

Nous donnons à la figure 3.3 la position des pôles et des zéros d'un filtre récursif qui approxime notre filtre passe-bas. L'erreur quadratique pour ce filtre (qui a 2 pôles et 13 zéros) est 0.0045, soit 35% de l'erreur précédente pour un même nombre d'opérations.

Chapitre 4

Échantillonnage des signaux

Dans ce chapitre nous répondons à la question : étant donné les échantillons $u_n = f(n)$ d'une fonction définie sur $\mathbb{R}$, peut-on reconstruire la fonction f à partir de la seule donnée de la suite définie sur $\mathbb{Z}$, u_n ?

Cette question se pose naturellement lorsque l'on veut stocker un signal sous forme numérique. Nous commençons par quelques exemples où la réponse est négative, puis nous arriverons à l'énnoncé du théorème de Shannon.

4.1 Exemples

premier exemple : Mélange de deux cosinus

Soit un signal défini sur $\mathbb{R}$ par (pour une raison ou pour une autre nous savons à son sujet qu'il est somme de deux fonctions cosinus de fréquences bien précises)

$$f(x) = \alpha \cos(2\pi x) + \beta \cos(2\pi 2x).$$

Connaître une telle fonction est équivalent à connaître les coefficients α et β. Supposons que l'on ne connaisse que les valeurs sur $\mathbb{Z}$ de ce signal, soit

$$\forall n \in \mathbb{Z},\ u_n = f(n) = \alpha \cos(2\pi n) + \beta \cos(2\pi 2n)$$

Comme n est un entier dans ce qui précède, on a

$$u_n = \alpha + \beta.$$

Ainsi, la connaissance des échantillons de la fonction f ne nous permet pas de connaître la fonction f. Les échantillons ne nous donnent accès qu'à la somme $\alpha+\beta$ et non à chacun des coefficients séparément. La raison en est que les deux fonctions

$$x \mapsto \cos(2\pi x)$$

et

$$x \mapsto \cos(2\pi 2x)$$

prennent les mêmes valeurs lorsque x est un entier. À tous les points d'échantillonnage, elles sont égales et ne peuvent être distinguées l'une de l'autre, les coefficients qui les

précèdent dans l'équation de f ne peuvent donc être séparés l'un de l'autre, si l'on ne connaît que les valeurs sur $\mathbb{Z}$ de f.

second exemple : Une onde pure

Soit maintenant la fonction à valeurs complexes définie par

$$g(x) = \alpha e^{2i\pi\lambda x}$$

où λ est un réel inconnu et α est un coefficient inconnu.

Connaître la fonction g revient à connaître α et λ. Supposons encore que l'on ne connaisse que les valeurs de g aux points entiers, soit :

$$\forall n \in \mathbb{Z},\ v_n = g(n) = \alpha e^{2i\pi\lambda n} = \alpha C^n$$

(où $C = e^{2i\pi\lambda}$ est une constante complexe de module 1.) Comment retrouver α et λ ?

Il est clair que $\alpha = v_0$.

Par contre, il y a une incertitude sur la valeur de λ. En effet, la seule condition sur λ imposée par la connaissance des échantillons v_n est

$$e^{2i\pi\lambda} = C = \frac{v_1}{v_0} \tag{4.1}$$

Comme C est de module 1, l'équation 4.1 a au moins une solution en λ. Mais cette solution n'est pas unique. On sait que si λ_0 est solution de l'équation 4.1 alors tous les réels de la forme

$$\lambda = \lambda_0 + n$$

où $n \in \mathbb{Z}$ sont les (seules) solutions de l'équation.

Remarques sur ces deux exemples

Dans les deux exemples précédents ce qui nous a empêché de reconstituer le signal à partir de ses échantillons c'est le fait que deux ondes pures de fréquences différentes prennent des valeurs égales aux points entiers. Plus précisément, les ondes pures sur $\mathbb{R}$ dont les fréquences diffèrent d'un entier (par exemple λ_0 et $\lambda_0 + 1$) sont indistinctes. En un certain sens, cela est la seule limite que pose l'échantillonnage à la connaissance complète du signal d'origine. Si l'on sait que les ondes pures contenues dans un signal (c.-à-d. support de sa transformée de Fourier) sont toutes dans un intervalle de longueur 1, alors la reconstruction sera possible.

Retour sur le premier exemple, avec des fréquences plus faibles

On va reprendre le premier exemple où l'on aura remplacé les fréquences 1 et 2 qui constituent la fonction f par les fréquences $\frac{1}{2}$ et 1, qui sont proches à moins de 1 près. Soit donc, f qui s'écrit :

$$f(x) = \alpha\cos(2\pi\frac{1}{2}x) + \beta\cos(2\pi x)$$

et que l'on dispose des échantillons aux points entiers

$$u_n = f(n) = \alpha\cos(\pi n) + \beta\cos(2\pi n) = (-1)^n\alpha + \beta$$

(car n est entier)

On a donc $u_0 = \alpha + \beta$ et $u_1 = -\alpha + \beta$. Soit encore,

$$\alpha = \frac{u_0 - u_1}{2} \text{ et } \beta = \frac{u_0 + u_1}{2}.$$

Ainsi, lorsque l'on sait que le signal ne porte que de "faibles" fréquences, on peut le connaître entièrement à partir de ses échantillons.

4.2 Formule de Poisson et théorème de Shannon

Nous allons aborder l'effet de l'échantillonnage sur le spectre d'une fonction. Soit f une fonction intégrable et $\hat{f}$ sa transformée de Fourier. Comme vu au chapitre précédent, $\hat{f}$ est définie sur $\mathbb{R}$. On considère la suite u définie par

$$u_n = f(n)$$

Sa transformée de Fourier (à temps discret) est notée $\hat{u}$ et est définie sur $[-\frac{1}{2}, \frac{1}{2}[$. Nous cherchons un moyen de décrire $\hat{u}$ à partir de $\hat{f}$.

raisonnement informel

- On l'a vu au chapitre précédent, $\hat{u}$ peut être vue comme une fonction périodique définie sur $\mathbb{R}$ (par périodisation de sa définition sur $[-\frac{1}{2}, \frac{1}{2}[$.
- Passer de $\hat{f}$ à f est une opération linéaire (transformée de Fourier inverse). Passer de f à u est une opération linéaire. Passer de u à $\hat{u}$ est aussi une opération linéaire (transformation de Fourier à temps discret. Donc, passer de $\hat{f}$ à $\hat{u}$ est une opération linéaire.
- Translater $\hat{f}$ de y revient à multiplier f par une onde pure ($t \mapsto e^{2i\pi yt}$). Multiplier f par une onde pure revient à multiplier u par une onde pure (définie sur $\mathbb{Z}$ et de fréquence y). Multiplier u par une onde pure revient à translater $\hat{u}$ (vue comme une fonction définie sur $\mathbb{R}$) dans les mêmes proportions que la translation initiale de $\hat{f}$.
- Nous en concluons que la relation qui lie $\hat{u}$ et $\hat{f}$ est une SLI. Il doit exister un noyau de convolution h qui permet d'écrire (pour tout f)

$$\hat{u} = \hat{f} * h$$

- C'est ici que les limites de notre raisonnement informel sont atteintes, car nous ne disposons pas dans ce cours des outils suffisants pour décrire h. En effet, h est ici une distribution (peigne de Dirac) et non une fonction. Néanmoins ce raisonnement nous permet de bien comprendre la raison de l'existence de la formule de Poisson que nous allons énoncer et aurait pu être complet si nous avions abordé des outils mathématiques plus complexes.

Comme vu dans les exemples introductifs, les ondes (sur $\mathbb{R}$) distantes d'un entier sont indistinctes une fois échantillonnées sur $\mathbb{Z}$. Il est naturel de considérer que les $\hat{f}(\nu + n)$ quand n parcourt $\mathbb{Z}$ jouent des rôles équivalents ce qui nous conduit à la formule qu'il faudrait prouver

$$\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[, \text{ (ou } \nu \in \mathbb{R}) \text{ , } \hat{u}(\nu) = \sum_{n \in \mathbb{Z}} \hat{f}(n + \nu) \tag{4.2}$$

Cette dernière équation est compatible avec la formule de convolution (à condition de remplacer l'intégration sur $\mathbb{R}$ par une somme discrète). En tout cas elle est bien linéaire et invariante par translation de $\hat{f}$. Elle signifie que lorsqu'une fonction est échantillonnée, la transformée de Fourier est périodisée.

Si nous l'écrivons en remplaçant $\hat{u}(\nu)$ par sa valeur en fonction de u (et donc en fonction de f).

$$\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[\sum_{m\in\mathbb{Z}} f(m)e^{2i\pi m\nu} = \sum_{n\in\mathbb{Z}} \hat{f}(n+\nu) \tag{4.3}$$

Il s'agit de la formule de Poisson, dont la principale conséquence est l'équation 4.2.

Théorème 4.1. Formule de Poisson ou le repliement spectral

Si f est une fonction définie sur $\mathbb{R}$ intégrable et telle que sa transformée de Fourier $\hat{f}$ est aussi intégrable et que la suite $u_m = f(m)$ est sommable et la suite $\hat{f}(n)$ est aussi sommable, alors on a

$$\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[\sum_{m\in\mathbb{Z}} f(m)e^{-2i\pi m\nu} = \sum_{n\in\mathbb{Z}} \hat{f}(n+\nu).$$

Par ailleurs, les formules

$$\sum_{n\in\mathbb{Z}} \hat{f}(n) = \sum_{m\in\mathbb{Z}} f(m)$$

et

$$\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[\hat{u}(\nu) = \sum_{n\in\mathbb{Z}} \hat{f}(n+\nu).$$

découlent de la formule de Poisson. La première en prenant $\nu = 0$ et la seconde est la constatation que le membre de gauche de l'équation de Poisson est l'expression de la TFtD de la suite des échantillons de f.

Remarque 4.2.
La formule de Poisson peut s'interpréter comme un repliement spectral, encore appelé aliasing. Ceci signifie que lorsque l'on échantillonne un signal son spectre subi une périodisation dont la formule

$$\nu \mapsto \sum_{n\in\mathbb{Z}} \hat{f}(n+\nu)$$

est l'expression mathématique.

Le mot anglais "aliasing" signifie littéralement "renommer" il reflète le fait que les fréquences $\nu + n$ sont vues comme équivalentes et "portent le même nom ν" (pour $\nu \in [-\frac{1}{2}, \frac{1}{2}[$).

Le mot français "repliement" vient du fait que le spectre est périodisé, mais aussi, comme on manipule le plus souvent des spectres de fonctions à valeurs réelles, ceux-ci sont symétriques (en module). Or, symétriser par rapport à 0 et translater de 1, revient à symétriser par rapport à 1/2.

Les figures 4.2 et 4.1 montrent graphiquement ce que signifie la périodisation du spectre.

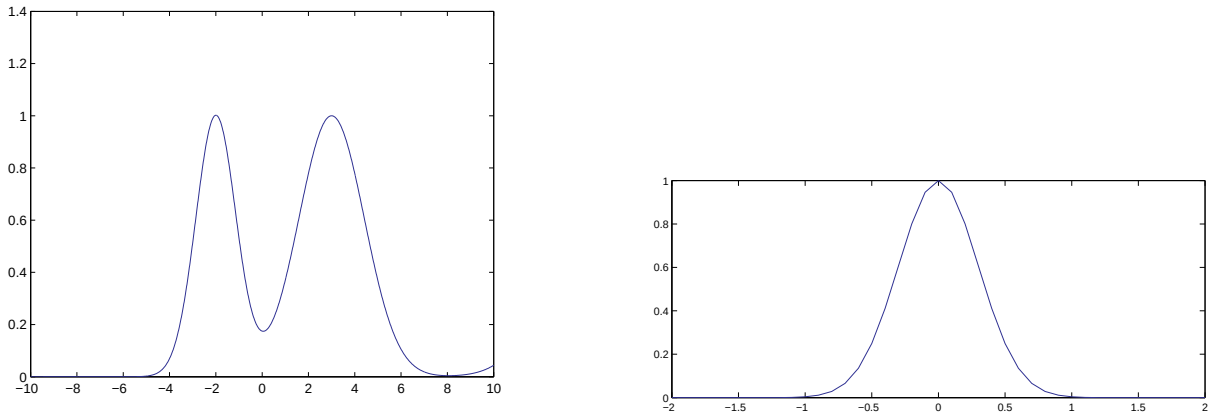


FIGURE 4.1 – Une fonction (à gauche) et son spectre (à droite)

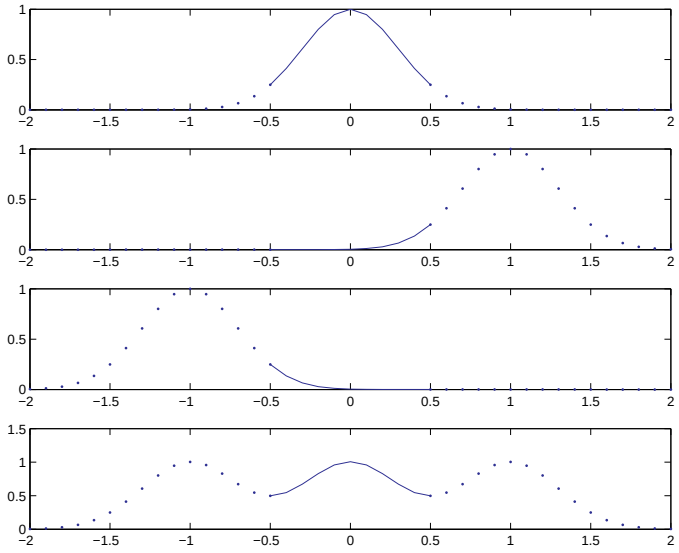


FIGURE 4.2 – Une explication graphique de la formule $\nu \mapsto \sum_{n\in\mathbb{Z}} \hat{f}(n+\nu)$. À la figure 4.1 on a vu une fonction et son spectre. En haut de la présente figure on a retracé le spectre en mettant en pointillé la partie hors de $[-\frac{1}{2}, \frac{1}{2}[$. Seconde ligne : la fonction $\nu \mapsto \hat{f}(\nu - 1)$ soit la décalé de 1 du spectre. Encore une fois on a mis en pointillé la partie hors de $[-\frac{1}{2}, \frac{1}{2}[$. troisième ligne $\nu \mapsto \hat{f}(\nu + 1)$ (translation de -1). Dernière ligne : La somme des trois translatées (0,1 et -1). Si on avait fait ce processus pour tous les $n \in \mathbb{Z}$ on aurait sur $[-\frac{1}{2}, \frac{1}{2}[$ la TFtD de la suite des échantillons $f(n)$. Ce qui est le sens de la formule de Poisson. On remarque que si le spectre avait été nul hors de $[-\frac{1}{2}, \frac{1}{2}[$ alors la somme totale aurait été strictement égale dans $[-\frac{1}{2}, \frac{1}{2}[$ au spectre d'origine. C'est cette remarque que le théorème de Shannon de bon échantillonnage exploite.

Démonstration 4.3. On considère la fonction g définie sur $[-\frac{1}{2},\frac{1}{2}[$ par

$$\forall \nu \in [-\frac{1}{2},\frac{1}{2}[, g(\nu) = \sum_{n\in\mathbb{Z}} \hat{f}(n+\nu).$$

Le théorème de Fubini permet de dire que g est intégrable sur $[-\frac{1}{2},\frac{1}{2}[$ (car $\hat{f}$ l'est sur $\mathbb{R}$).

On veut calculer les coefficients de Fourier de g. On les note c_m et ils sont définis sur $\mathbb{Z}$ par

$$\begin{aligned}
\forall m \in \mathbb{Z}, c_m &= \int_{\nu=-\frac{1}{2}}^{\frac{1}{2}} g(\nu) e^{-2i\pi m\nu} d\nu \\
&= \int_{\nu=-\frac{1}{2}}^{\frac{1}{2}} \left[\sum_{n\in\mathbb{Z}} \hat{f}(n+\nu)\right] e^{-2i\pi m\nu} d\nu \\
&= \int_{\nu=-\frac{1}{2}}^{\frac{1}{2}} \left[\sum_{n\in\mathbb{Z}} \hat{f}(n+\nu) e^{-2i\pi m\nu}\right] d\nu \\
&= \int_{\nu=-\frac{1}{2}}^{\frac{1}{2}} \left[\sum_{n\in\mathbb{Z}} \hat{f}(n+\nu) e^{-2i\pi m(\nu+n)}\right] d\nu \text{ car } e^{-2i\pi mn} = 1 \\
&= \int_{z\in\mathbb{R}} \hat{f}(z) e^{-2i\pi mz} dz \text{ en posant } z = n+\nu \text{ et en remarquant que } z \text{ parcourt } \mathbb{R}. \\
&= f(-m) \text{ par le théorème d'inversion}
\end{aligned}$$

Comme la suite $f(-m)$ est sommable, par hypothèse, on peut appliquer le théorème d'inversion à la fonction g. Ce qui s'écrit

$$\forall \nu \in [-\frac{1}{2},\frac{1}{2}[,\ g(\nu) = \sum_{m\in\mathbb{Z}} f(-m) e^{2i\pi\nu m} = \sum_{m\in\mathbb{Z}} f(m) e^{-2i\pi\nu m}$$

En remplaçant $g(\nu)$ par sa valeur en fonction de $\hat{f}$ on a

$$\forall \nu \in [-\frac{1}{2},\frac{1}{2}[,\ \sum_{n\in\mathbb{Z}} \hat{f}(n+\nu) = \sum_{m\in\mathbb{Z}} f(m) e^{-2i\pi\nu m}$$

Théorème 4.4. Théorème de bon échantillonnage, ou théorème de Shannon *Si f est une fonction sommable et que sa TFtC, $\hat{f}$, est nulle hors de l'intervalle $[-\frac{1}{2},\frac{1}{2}]$. Comme $\hat{f}$ est bornée (par $\|f\|_1$) elle est donc sommable. On suppose de plus que les $f(n)$ (pour n entier) est une suite sommable. Alors on a*

$$\forall t \in \mathbb{R},\ f(t) = \sum_{n\in\mathbb{Z}} f(n)\,\mathrm{sinC}(\pi(t-n))$$

et de plus la TFtD de la suite des échantillons $f(m)$ de f est $\hat{f}$ (qui est nulle hors de $[-\frac{1}{2},\frac{1}{2}[$) c'est-à-dire :

$$\forall \nu \in [-\frac{1}{2},\frac{1}{2}[,\ \sum_{m\in\mathbb{Z}} f(m) e^{2i\pi\nu m} = \hat{f}(\nu)$$

En particulier, deux fonctions $f, g \in L^1(\mathbb{R})$ qui ont leur TF à support dans $[-\frac{1}{2}, \frac{1}{2}]$ et vérifient les autres hypothèses de ce théorème alors

$$(\forall n \in \mathbb{Z},\ f(n) = g(n) \quad) \Longrightarrow (\ f = g\ )$$

i.e. l'opération d'échantillonnage sur $\mathbb{Z}$ est injective sur l'espace des fonctions dont la TF est à support dans $[-\frac{1}{2}, \frac{1}{2}[$.

Démonstration 4.5.

D'abord, on remarque que $\hat{f}$ est une fonction continue, comme elle est nulle hors de $[-\frac{1}{2}, \frac{1}{2}]$ on a $\hat{f}(-\frac{1}{2}) = \hat{f}(\frac{1}{2}) = 0$.

La fonction f vérifie les hypothèses de l'équation de Poisson et on a donc

$$\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[,\ \sum_{m\in\mathbb{Z}} f(m)e^{-2i\pi\nu n} = \sum_{n\in\mathbb{Z}} \hat{f}(\nu + n)$$

Mais le terme de droite de la dernière égalité est nul pour $n \neq 0$, on a donc

$$\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[,\ \sum_{m\in\mathbb{Z}} f(m)e^{-2i\pi\nu m} = \hat{f}(\nu)$$

Comme $\hat{f}$ est sommable on peut appliquer le théorème d'inversion et on a

$$\forall t \in \mathbb{R},\ f(t) = \int \hat{f}(\nu)e^{2i\pi t\nu}d\nu = \int_{-\frac{1}{2}}^{\frac{1}{2}} \hat{f}(\nu)e^{2i\pi t\nu}$$

$$= \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{m\in\mathbb{Z}} f(m)e^{2i\pi\nu(t-m)}d\nu = \sum_{m\in\mathbb{Z}} f(m) \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2i\pi\nu(t-m)}d\nu = \sum_{m\in\mathbb{Z}} f(m)\,\mathrm{sinC}(\pi(t-m))$$

L'injéctivité découle de la dernière formule puisque f (et g) ont les mêmes échantillons et que cette formule donne la valeur de $f(t)$ (et $g(t)$) à partir de la valeur des échantillons.

Théorème 4.6. Théorème de Shannon pour les fonctions d'énergie finie
Si f est une fonction d'énergie finie ($f \in L^2(\mathbb{R})$) telle que son spectre est à support dans $[-1/2, 1/2]$. On note $u_n = f(n)$ alors on a

$$\|f\|_2 = \|u\|_2$$

Autrement dit, la norme de la suite des échantillons de f est la même que celle de f. De cela il découle que si f et g vérifient l'hypothèse sur le spectre et que

$$\forall n \in \mathbb{Z}, f(n) = g(n)$$

alors

$$f = g$$

(l'égalité des échantillons implique l'égalité des fonctions) Pour voir cela il suffit de remarquer que $f - g$ vérifie les hypothèses et que ses échantillons sont tous nuls ce qui implique que

$$\|f - g\|_2 = 0.$$

Par ailleurs on a

$$f = \sum_{n\in\mathbb{Z}} f(n)\,\mathrm{sinC}_n^\pi\,.$$

où $\mathrm{sinC}_n^\pi(x) = \mathrm{sinC}(\pi(x-n))$*. Cette égalité est à prendre au sens* L^2 *il s'agit d'une somme infinie de fonctions* L^2 *(les* sinC_n^π*) affectées de coefficient* l^2 *(les échantillons de* f*). Naïvement on pourrait écrire*

$$\forall t, f(t) = \sum_{n\in\mathbb{Z}} f(n)\,\mathrm{sinC}(\pi(t-n))$$

mais cette égalité ponctuelle n'est vraie que si l'on a vérifié que les échantillons $f(n)$ *son sommables* $((f(n))_n \in l^1)$*.*

Remarque 4.7. **Cas d'une fonction** L^1
Si f est sommable et que sa TF est à support borné, alors $\hat{f}$ est sommable (car continue et de support borné). $\hat{f}$ est même d'énergie finie. Le théorème d'inversion s'applique et l'égalité de Parseval nous dit que f est elle aussi d'énergie finie. Et donc, f rentre dans le cadre du théorème ci-dessus.

Démonstration 4.8.
Comme $\hat{f}$ est L^2 et qu'elle est à support compact, elle est aussi L^1 et f est donc une fonction continue et la valeur de ses échantillons (valeurs ponctuelles) a un sens.

On considère la fonction g définie sur $[-1/2, 1/2]$ par

$$\forall \nu \in [-1/2, 1/2], g(\nu) = \sum_{n\in\mathbb{Z}} \hat{f}(n+\nu) = \hat{f}(\nu)$$

La deuxième égalité est due au fait que le spectre de f est à support dans $[-1/2, 1/2]$. Comme dans la démonstration de la formule de Poisson, la fonction g a pour transformée de Fourier (coefficients de Fourier dans ce cas) la suite des échantillons de f, c'est-à-dire la suite u. Or l'égalité de Parseval nous dit que

$$\|g\|_2 = \|u\|_2$$

Par ailleurs l'égalité

$$\forall \nu \in [-1/2, 1/2], g(\nu) = \hat{f}(\nu)$$

nous dit que

$$\|g\|_2 = \|\hat{f}\|_2 = \|f\|_2$$

ce qui prouve le théorème. La seconde partie (injectivité de l'échantillonnage) est déjà expliquée dans le corps du théorème.

Preuve de l'égalité (dans L^2)

$$f = \sum_{n\in\mathbb{Z}} f(n)\,\mathrm{sinC}_n^\pi\,.$$

Pour voir cela, il suffit de constater que les fonctions

$$f_N = f - \sum_{n=-N}^{n=N} f(n)\,\mathrm{sinC}_n^\pi$$

Vérifient les hypothèses du présent théorème (sont bien L^2 et ont leur spectre dans $[-\frac{1}{2}, \frac{1}{2}]$). Elles ont donc une norme L^2 égale à celle de leurs échantillons, soit

$$\int |f_N|^2 = \sum_{|n|>N} |f(n)|^2$$

Cette dernière somme tend vers 0 lorsque N tend vers l'infini, ce qui signifie que la suite de fonctions f_N tend vers la fonction nulle en norme L^2 et donc que la suite de fonctions

$$\sum_{n=-N}^{n=N} f(n)\,\mathrm{sinC}_n^\pi$$

tend vers f, ce qu'il fallait démontrer.

4.3 Reconstruction

4.3.1 Reconstruction parfaite

Les formules

$$\forall t \in \mathbb{R},\ f(t) = \sum_{n\in\mathbb{Z}} f(n)\,\mathrm{sinC}(\pi(t-n))$$

et (plus généralement dans le cas L^2)

$$f = \sum_{n\in\mathbb{Z}} f(n)\,\mathrm{sinC}_n^\pi$$

Signifient que si le spectre de f est à support dans $[-\frac{1}{2}, \frac{1}{2}]$ alors on peut reconstruire la fonction f à partir de ses échantillons. On appelle une telle reconstruction, la reconstruction parfaite. L'appareil qui, partant de la suite des échantillons $f(n)$ produit la somme

$$\forall t \in \mathbb{R},\ f(t) = \sum_{n\in\mathbb{Z}} f(n)\,\mathrm{sinC}(\pi(t-n))$$

s'appelle un **CNA parfait** ou **CNA idéal**. (CNA=Convertisseur Numérique Analogique)

Les figures 4.3 et 4.4 montrent comment on reconstruit une fonction à partir de ses échantillons par cette méthode. Le Théorème de Shannon dit que si f est à bande limité dans $[-\frac{1}{2}, \frac{1}{2}]$ alors cette reconstruction redonne la fonction d'origine. Il s'agit d'une réponse complète à la question que nous nous sommes posée en début de chapitre.

4.3.2 Autres reconstructions

Dans la pratique il est impossible d'avoir un CNA parfait car, en particulier, la fonction sinC est à support infini et un tel appareil aurrait besoin de manipuler une infinité d'échantillons du signal pour pouvoir calculer une seule valeur $f(t)$.

Nous présentons ici des types de reconstructeurs. Il en existe d'autres qui sont de plus en plus complexes et produisent des approximations meilleurs de la fonction d'origine.

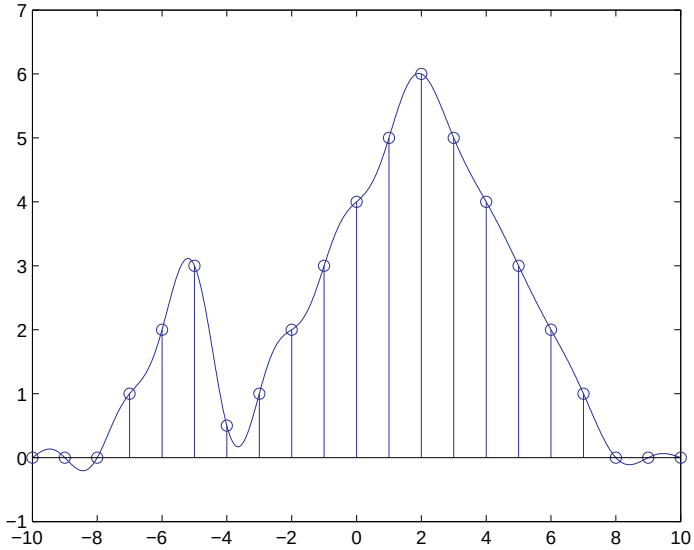


FIGURE 4.3 – Une fonction qui vérifie les hypothèses du théorème de Shannon. Les échantillons sont mis en évidence. On montre à la figure 4.4 comment la formule de reconstruction sert à reconstruire cette fonction à partir de ses échantillons.

Reconstructeur d'ordre 0 ou Bloqueur

Un bloqueur d'ordre 0 est un appareil qui reçoit en entrée une suite d'échantillons u_n et renvoie la fonction définie par

$$g(t) = u_n \text{ si } n \leqslant t < n+1$$

Autrement dit, le signal analogique en sortie est constant par morceaux. Entre les instants n et $n+1$ il vaut la valeur du dernier échantillon lu. D'où le nom de bloqueur. Si les $u_n = f(n)$ sont les échantillons de f on peut exprimer g par

$$g(t) = \sum f(n)\mathbb{1}_{[0,1[}(t-n)$$

La figure 4.5 montre la sortie associée par un bloqueur.

Reconstructeur d'ordre 1

La dernière reconstruction (ordre 0) donnait une fonction non continue. On sait que lorsqu'une une fonction vérifie les hypothèses de Shannon, elle est très régulière (La décroissance rapide du spectre à l'infini se traduite par une très grande régularité de la fonction).

Le reconstructeur d'ordre 1, renvoie une fonction continue mais non dérivable (aux point entiers).

Si u_n est la suite d'échantillons en entrée, le signal analogique en sortie a la valeur suivante

$$g(t) = (t-n)u_{n+1} + (1-(t-n))u_n \text{ si } n \leqslant t \leqslant n+1$$

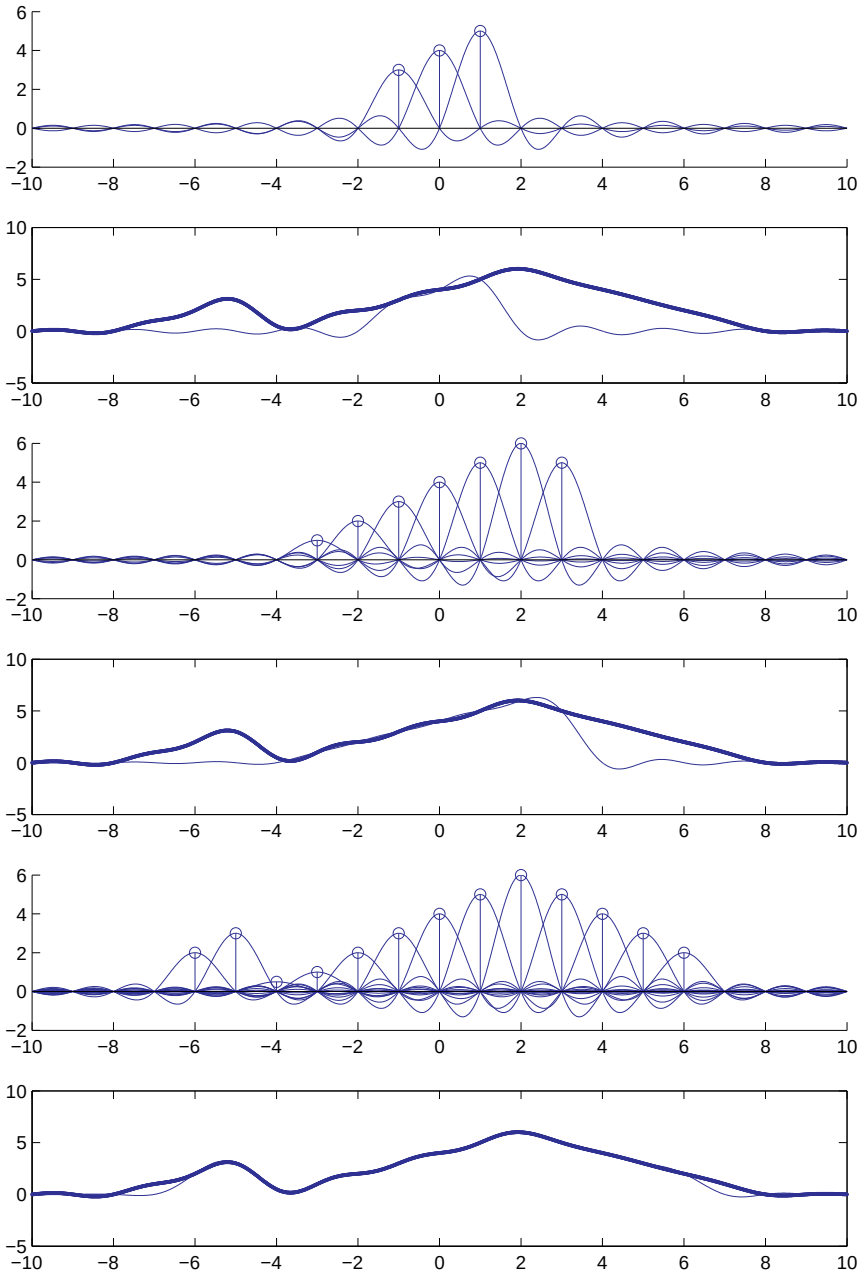


FIGURE 4.4 – Reconstruction d'une fonction à partir de ses échantillons par la formule d'interpolation de Shannon (en utilisant des fonctions sinC^π). À chaque fois on trace pour un certain nombre d'échantillons les fonctions $t \mapsto f(n)\,\mathrm{sinC}_n^\pi(t)$ et en dessous on trace la somme en trait fin. En trait gras il y a la fonction d'origine, celle de la figure 4.3. Plus on fait intervenir d'échantillons (3 puis 7 puis 13) plus on s'approche de la fonction d'origine (si celle-ci vérifie les hypothèses du théorème de Shannon).

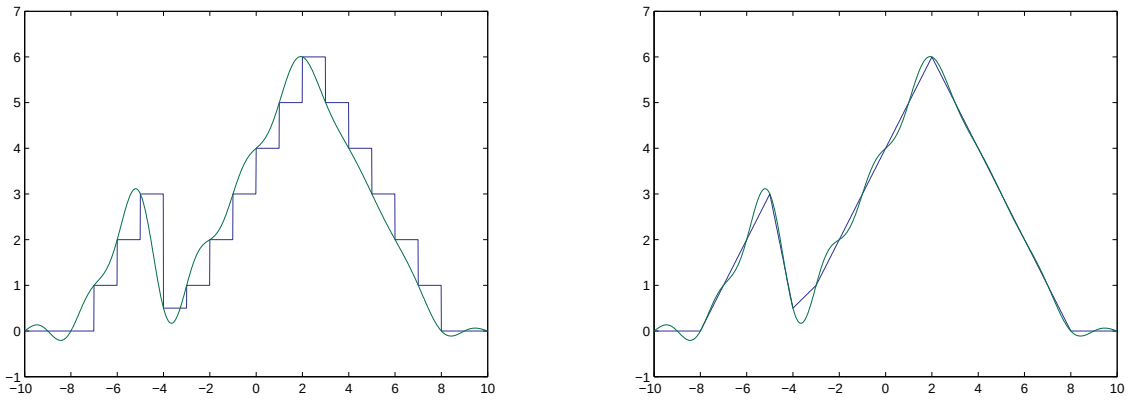


FIGURE 4.5 – Pour une même suite d'échantillons, nous montrons la sortie d'un bloqueur (à gauche) et d'un reconstructeur d'ordre 1 (à droite). À chaque fois nous avons surimposé la fonction d'origine. L'approximation à l'ordre 1 est évidemment meilleure que celle à l'ordre 0. Cependant, on peut montrer qu'une fonction qui vérifie le théorème de Shannon ne peut pas être strictement linéaire sur un intervalle.

Il s'agit d'une interpolation linéaire (polynôme de degré 1) entre les échantillons.

Si les $u_n = f(n)$ sont les échantillons d'une fonction f, on peut exprimer g par

$$g(t) = \sum f(n) h_1(t-n)$$

avec h_1 la fonction "triangle" définie par

$$h_1(t) = \begin{cases} 1 - |t| \text{ si } -1 \leqslant t \leqslant 1 \\ 0 \text{ sinon.} \end{cases}$$

La figure 4.5 montre la sortie associée par un reconstructeur d'ordre 1.

4.3.3 Erreur quadratique de reconstruction

On veut savoir à quel point la fonction reconstruite à partir des échantillons est une bonne approximation de la fonction d'origine.

Soit f une fonction qui vérifie les hypothèses du théorème de l'équation de Poisson. Soit $u_n = f(n)$ la suite de ses échantillons. On appelle g la sortie associée par un reconstructeur à la suite des échantillons. On suppose que le reconstructeur agit de la manière suivante (ce qui est le cas de ceux vus ici)

$$\forall t \in \mathbb{R},\ g(t) = \sum_n u_n h(t-n) = \sum_n f(n) h(t-n)$$

où h est la fonction de reconstruction (c'est le créneau pour le reconstructeur d'ordre 0 et la fonction triangle pour le reconstructeur d'ordre 1). Et on se demande quelle est l'erreur de reconstruction notée E_q (en fonction de h et f) qui est commise. On veut évaluer

$$E_q = \|f - g\|_2$$

Par l'égalité de Parseval on a

$$E_q = \|\hat{f} - \hat{g}\|_2$$

Quel est le spectre de g ? D'après les propriétés usuelles de la TF, la TFtC de la fonction

$$t \mapsto f(n)h(t-n)$$

(c'est la fonction h translatée de n et multipliée par la constante $f(n)$) est

$$\nu \mapsto \hat{h}(\nu)e^{-2i\pi n\nu}f(n)$$

On en déduit que le spectre de g est

$$\hat{g}(\nu) = \hat{h}(\nu)\sum_{n\in\mathbb{Z}} f(n)e^{-2i\pi\nu n} = \hat{h}(\nu)\sum_{n\in\mathbb{Z}}\hat{f}(n+\nu)$$

La dernière égalité étant l'application de l'équation de Poisson. Nous allons expliciter l'erreur quadratique de reconstruction dans **le cas où la fonction f a son spectre supporté par** $[-\frac{1}{2},\frac{1}{2}]$ (c'est à dire dans le cadre d'un bon échantillonnage). Dans ce cas le spectre de g est le périodisé du spectre de f multiplié par le spectre de h.

$$\hat{g}(\nu) = \hat{h}(\nu).f(\nu_0) \text{ où } \nu_0 \text{ vérifie } \nu_0 \in [-\frac{1}{2},\frac{1}{2}[\text{ et } \nu - \nu_0 \in \mathbb{Z}$$

On pose le calcul de E_q^2

$$E_q^2 = \int_{\nu\in\mathbb{R}} |\hat{f}(\nu) - \hat{g}(\nu)|^2 d\nu = \sum_{n\in\mathbb{Z}}\int_{n-\frac{1}{2}}^{n+\frac{1}{2}} |\hat{f}(\nu) - \hat{g}(\nu)|^2 d\nu$$

On distingue le cas $n = 0$ (car f a son spectre dans $[-\frac{1}{2},\frac{1}{2}]$) des autres et on a

$$E_q^2 = \int_{-\frac{1}{2}}^{\frac{1}{2}} |\hat{f}(\nu)|^2.|1-\hat{h}(\nu)|^2 d\nu + \int_{-\frac{1}{2}}^{\frac{1}{2}} |\hat{f}(\nu)|^2 \sum_{n\neq 0}|\hat{h}(\nu+n)|^2$$

La première partie de la somme est d'autant plus petite que $\hat{h}$ est proche de 1 sur $[-\frac{1}{2},\frac{1}{2}]$. La seconde est d'autant plus petite que $\hat{h}$ est petit hors de $[-\frac{1}{2},\frac{1}{2}]$. Et, évidemment, E_q est nulle si $h = \text{sinC}$.

La meilleure reconstruction possible ? On se pose le problème suivant : étant donné une fonction f, quelles valeurs u_n introduire dans un CNA idéal (celui qui reconstruit avec des sinC^π) afin que sa sortie, notée g et définie par

$$g(t) = \sum_n u_n \,\text{sinC}(\pi(t-n))$$

soit la plus proche possible de la fonction f, c.-à-d. que $E_q = \|f-g\|_2$ soit minimale ?

Appelons $f_B = f * \text{sinC}^\pi$, alors f_B vérifie les hypothèses du théorème de Shannon. Appelons $f_H = f - f_B$.

On a

$$\forall g = \sum u_n \,\text{sinC}_n^\pi,\ \|f-g\|_2^2 = \|f_H + (f_B - g)\|_2^2$$

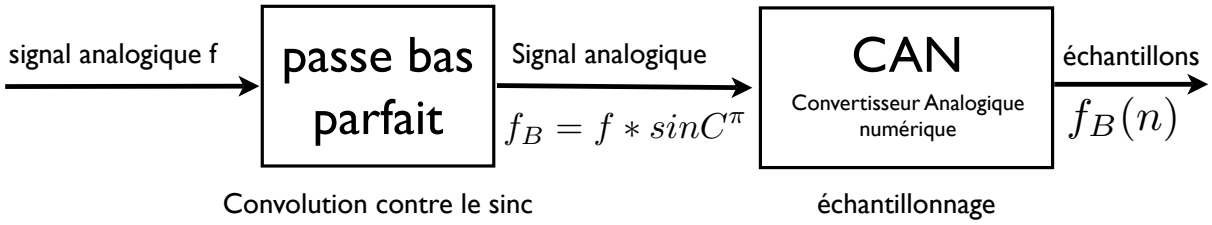


FIGURE 4.6 – Chaîne de bon échantillonnage. Avant d'échantillonner un signal, et afin d'éviter le repliement spectral, il convient d'appliquer un passe bas parfait qui coupe à partir de la fréquence $F_e/2$ (soit $1/2$ si on échantillonne sur $\mathbb{Z}$, voir la section normalisation)

Or f_H a son spectre supporté hors de $[-\frac{1}{2}, \frac{1}{2}]$ elle est donc orthogonale à $f_B - g$ pour tout choix de u_n (on exprime le produit scalaire comme la valeur en 0 d'une convolution, et d'après les règles de la TFtC, une convolution est nulle si le produit des TFtC est nul). Par le théorème de Pythagore

$$E_q^2 = \|f_H\|_2^2 + \|f_B - g\|_2^2$$

D'un autre coté, si on prend pour $u_n = f_B(n)$ on a $\|f_B - g\|_2 = 0$ (par le théorème de Shannon).

Au final, la meilleure approximation L^2 est obtenue en choisissant pour les u_n les valeurs des échantillons de $f * \mathrm{sinC}^\pi$

$$u_n = (f * \mathrm{sinC}^\pi)(n)$$

4.4 Chaîne d'échantillonnage

Dans la pratique, aucun signal réel ne peut être à bande strictement limitée. Il faut donc, avant d'échantillonner un signal, le filtrer afin d'éliminer les hautes fréquences et éviter le phénomène de repliement spectral.

Pour cela il faut faire passer le signal par un SLI passe-bas parfait qui coupe à la fréquence $\frac{1}{2}$. Cela revient à convoler le signal avec la fonction sinC^π. Ainsi, même si la reconstruction n'est pas strictement égale au signal d'origine (nous avons perdu la partie haute fréquence notée f_H ci-dessus) il reste qu'il n'y aura pas de parasitage du spectre de f_B par des composantes de f_H.

La figure 4.6 montre la chaîne d'échantillonnage parfait complète.

4.5 Normalisation

Nous avons, pour simplifier la présentation, traité seulement le cas où l'échantillonnage s'est fait sur $\mathbb{Z}$. Que ce passe-t-il si l'on veut échantillonner un signal à une fréquence différente de la fréquence 1 ?

Soit f une fonction et $F_e > 0$ un réel. On appel F_e la fréquence d'échantillonnage et $T_e = \frac{1}{F_e}$ est appelé **période d'échantillonnage**.

On se demande si on peut reconstruire f à partir de la suite des échantillons $f(n.T_e)$?

On a vu que le spectre de la fonction g définie par

$$g(x) = f(xT_e) = f(\frac{x}{F_e})$$

est donné par

$$\hat{g}(\nu) = \frac{1}{T_e}\hat{f}(\frac{1}{T_e}\nu) = F_e\hat{f}(F_e\nu)$$

et

$$\hat{f}(\nu) = T_e\hat{g}(\frac{\nu}{F_e})$$

Or, poser la question de la reconstruction de f à partir des $f(nT_e)$ est équivalent à vouloir reconstruire g à partir des $g(n)$ ($n \in \mathbb{Z}$). Par exemple, il faut, pour satisfaire aux conditions du théorème de Shannon que le spectre g soit à support dans $[-\frac{1}{2}, \frac{1}{2}]$. Cela se traduit sur le spectre de f comme ceci

$$(\forall|\nu| > \frac{1}{2}\hat{g}(\nu) = 0) \Leftrightarrow (\forall|\xi| > \frac{F_e}{2}\hat{f}(\xi) = 0)$$

Après ces remarques les équations de Poisson et la condition du théorème de Shannon (les conditions d'intégrabilité, sommabilité, énergie finie restent les mêmes)

4.5.1 Formule de Poisson à la fréquence F_e

Sous les hypothèses d'intégrabilité adéquates on a

$$\forall\nu \in [-\frac{1}{2}, \frac{1}{2}[\sum_{m\in\mathbb{Z}} f(mT_e)e^{-2i\pi m\nu} = F_e\sum_{n\in\mathbb{Z}}\hat{f}(F_e(n+\nu)) = \frac{1}{T_e}\sum_{n\in\mathbb{Z}}\hat{f}(F_e(n+\nu))$$

Soit, la TFtD de la suite $f(nT_e)$ est la périodisée à F_e du spectre de f (que l'on normalise pour faire coïncider $\nu = 1/2$ avec la fréquence $Fe/2$) (sans oublier la multiplication par $1/T_e$).

4.5.2 Théorème de Shannon

Si f a son spectre supporté par $[-F_e/2, F_e/2]$ alors f peut-être reconstruite grâce aux échantillons $f(nT_e)$ par la formule

$$f(t) = \sum_n f(nT_e)\,\mathrm{sinC}\left(\pi(\frac{t}{T_e} - n)\right) = \sum_n f(nT_e)\,\mathrm{sinC}\left(\frac{\pi}{T_e}(t - nT_e)\right)$$

4.5.3 Chaîne d'échantillonnage

Dans la chaîne d'échantillonnage il faut passer le signal dans un filtre passe-bas parfait qui coupe à partir de la fréquence $F_e/2$ (et non plus 1/2). Il faut donc convoluer le signal par la fonction

$$t \mapsto F_e\,\mathrm{sinC}(\pi\frac{t}{T_e})$$

Vous pouvez retenir de manière mnémotechnique cette formule en disant que la valeur en 0 de cette fonction doit être égale à l'intégrale de $\mathbb{1}_{[-F_e/2,F_e/2]}$, c'est à dire F_e et que le sinus cardinal adéquat s'annule aux points d'échantillonnage (nT_e) sauf en 0.

Chapitre 5

Transformée en cosinus discret (DCT)

Dans ce court chapitre nous introduisons une la transformation en cosinus discret (Discrete Cosine Transform en anglais) qui est plus adaptée à l'étude des signaux finis que la TFD mais qui lui est très liée.

5.1 Définition et propriétés

Définition 5.1. *Soit $u_0, \dots, u_{N-1}$ un signal fini, on définit la transformée en cosinus discret de u, que l'on note $\hat{u}^D$ par*

$$\forall 0 \leqslant k \leqslant N-1,\ \hat{u}_k^D = w_k \sum_{n=0}^{N-1} u_n \cos\left(2\pi(n+\frac{1}{2})\frac{k}{2N}\right)$$

avec $w_0 = \sqrt{\frac{1}{N}}$ et $w_k = \sqrt{\frac{2}{N}}$ pour $k \neq 0$.

Cette définition ressemble à celle de la TFD à quelques différences près

1. L'exponentielle complexe est remplacée par un cosinus.
2. La position de l'échantillon n est remplacée par $n + 1/2$.
3. La fréquence k/N dans la TFD est remplacée $k/2N$.
4. Un facteur de normalisation qui diffère entre le cas $k = 0$ et $k \neq 0$ est appliqué, il a pour but de rendre orthonormée la base de la DCT.

Proposition 5.2. Lien avec la TFD
Si on définit le signal fini de taille $2N$, x par

$$x_n = \begin{cases} u_n \text{ si } n < N \\ u_{2N-1-n} \text{ si } N \leqslant n \leqslant 2N-1 \end{cases}$$

Autrement dit le signal x est la concaténation du signal u avec son symétrique.
Si on note $\hat{x}$ la TFD (d'ordre 2N) de x. Alors on a

$$\hat{u}_0^D = \frac{1}{2\sqrt{N}}\hat{x}_0$$

et

$$\hat{u}_k^D = e^{-i\pi\frac{k}{2N}}\frac{1}{\sqrt{2N}}\hat{x}_k \text{ pour } 1 \leqslant k \leqslant N-1 \tag{5.1}$$

Et de plus on a, si u est un signal réel,

$$\sum_k \left|\hat{u}_k^D\right|^2 = \sum_n |u_n|^2$$

Remarque : Cette proposition permet d'effectuer le calcul de la DCT de manière rapide à l'aide de l'algorithme de la FFT. En effet, pour calculer la DCT d'un signal il suffit de doubler sa taille en le symétrisant, puis de calculer la TFD du signal double. Enfin, on applique la formule 5.1.

Démonstration. On commence par montrer la formule 5.1 (le cas $k = 0$ est trivial). On se place dans le cas $k \neq 0$ et on calcule $\hat{x}_k$ la TFD (d'ordre $2N$) du signal x.

$$\begin{aligned}\hat{x}_k &= \sum_{n=0}^{2N-1} x_n e^{-2i\pi\frac{k}{2N}n} = \sum_{n=0}^{N-1} u_n e^{-2i\pi\frac{k}{2N}n} + \sum_{n=N}^{2N-1} u_{2N-1-n} e^{-2i\pi\frac{k}{2N}n} \\ &= \sum_{n=0}^{N-1} u_n e^{-2i\pi\frac{k}{2N}n} + \sum_{m=0}^{N-1} u_m e^{-2i\pi\frac{k}{2N}(2N-1-m)}\end{aligned}$$

(on a fait le changement de variable $m = 2N - 1 - n$ dans la seconde somme)

$$= \sum_{n=0}^{N-1} u_n \left(e^{-2i\pi\frac{k}{2N}n} + e^{-2i\pi\frac{k}{2N}(-1-n)}\right) = \sum_{n=0}^{N-1} u_n e^{i\pi\frac{k}{2N}} \left(e^{-2i\pi\frac{k}{2N}(n+1/2)} + e^{2i\pi\frac{k}{2N}(n+1/2)}\right)$$

$$= 2e^{i\pi\frac{k}{2N}} \cdot \sqrt{\frac{N}{2}} \hat{u}_k^D$$

D'où la formule 5.1.

Passons à la démonstration de

$$\sum_k \left|\hat{u}_k^D\right|^2 = \sum_n |u_n|^2$$

On sait par les propriétés de la TFD que

$$\sum_{k=0}^{2N-1} |\hat{x}_k|^2 = 2N \sum_{n=0}^{2N-1} |x_n|^2$$

Or, par définition de x,

$$\sum_{n=0}^{2N-1} |x_n|^2 = 2\sum |u_n|^2$$

car x est la répétition de u. Et donc,

$$\sum_{k=0}^{2N-1} |\hat{x}_k|^2 = 4N \sum_{n=0}^{N-1} |u_n|^2$$

Par ailleurs, comme le signal x est réel, on, pour tout $0 < k < N$

$$\hat{x}_k = \overline{\hat{x}_{2N-k}}$$

(symétrie hermitienne) et de plus, $\hat{x}_N = 0$ (par symétrie du signal x).

On a donc,

$$4N \sum_{n=0}^{2N-1} |u_n|^2 = \sum_{k=0}^{2N-1} |\hat{x}_k|^2 = |\hat{x}_0|^2 + 2 \sum_{k=1}^{N-1} |\hat{x}_k|^2$$

Enfin, si on remplace les $\hat{x}_k$ par leurs valeurs en fonction des $\hat{u}_k^D$

$$4N \sum_{n=0}^{2N-1} |u_n|^2 = 4N|\hat{u}_0^D|^2 + 2 \sum_{k=1}^{N-1} 2N|\hat{u}_k^D|^2 = 4N \sum |\hat{u}_k^D|^2$$

□

Définition 5.3. Base de la DCT *C'est la base orthonormée de $\mathbb{R}^N$ constituée des vecteurs indéxés par $k = 0 \dots N-1$ et de formule générale (terme numéro n) :*

$$n \mapsto w_k \cos\left(2\pi \left(n + \frac{1}{2}\right) \frac{k}{2N}\right)$$

Obtenir la DCT d'un signal c'est effectuer le produit scalaire contre ces vecteurs.

Cette base est orthonormée car

$$\sum_k |\hat{u}_k^D|^2 = \sum_n |u_n|^2$$

Définition 5.4. DCT locale
Pour les signaux de taille mN on appelle base de la DCT locale de taille N, l'ensemble des mN vecteurs obtenus en translatant les vecteurs de la base de la DCT de taille N aux positions multiples de N. (voir TD 4 exercice 1)

Définition 5.5. DCT2D
La base de la DCT bidimensionnelle de $\mathbb{R}^{N \times N}$ est celle obtenue en opérant le produit tensoriel sur la base de la DCT monodimensionnelle de taille N. Elle compte N^2 vecteurs. (voir TD 4 exercice 2)

Définition 5.6. DCT locale 2D
La DCT locale de taille $N \times N$ pour une image de taille $(mN) \times (mN)$ est la base que l'on obtient en décalant la base de la DCT 2D de taille $N \times N$ à toutes les positions multiples de N (dans les deux dimensions)

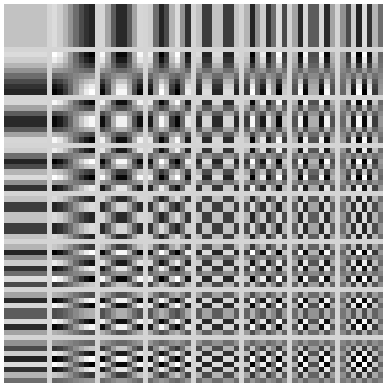


FIGURE 5.1 – Tous les éléments de la base de la DCT de taille 8×8. Il y en a 64. La fréquence en x croit de gauche à droite et la fréquence en y croit de haut en bas.

5.2 Décroissance des coefficients de la DCT comparée à celle de la TFD

On sait que la TFD d'un signal ne change pas en module si on translate circulairement un signal. Par exemple, le signal (1,2,3,4,5) aura des coefficients de TFD égaux en module à ceux du signal (5,1,2,3,4) car la translation circulaire se transforme en multiplication par une onde (de module 1) dans le domaine de Fourier.

Cela signifie que le dernier élément d'un signal fini et son premier son considérés comme proches lors du calcul de la TFD (autant que le premier et second élément). Évidemment, les signaux naturels même quand ils sont réguliers n'ont aucune raison d'avoir des valeurs proches entre la position 0 et la position $N-1$. Cela se traduit par une discontinuité qui se reflète en une décroissance lente des coefficients de la TFD, même lorsque le signal semble régulier.

Nous allons, sur un exemple particulier, montrer pourquoi la DCT ne présente pas ce problème. Nous considérons les deux signaux définis sur [-1/2,1/2[par

$$f(x) = x$$
$$g(x) = 2x + 1/2 \text{ si } x \leqslant 0 \text{ et } g(x) = -2x + 1/2 \text{ si } x \geqslant 0$$

Le signal g est le symétrisé de f.

On note c_n les coefficients de Fourier de f et d_n les coefficients de Fourier de g. La comparaison des suites c_n et d_n permet d'évaluer qualitativement la décroissance des coefficients d'une TFD par rapport à ceux d'une DCT, car la DCT revient aussi à symétriser un signal avant d'en calculer la TFD.

On a

$$c_n = \int_{-\frac{1}{2}}^{\frac{1}{2}} x e^{-2i\pi nx} dx = \frac{i(-1)^n}{2\pi n} \text{ et } (c_0 = 0)$$

et

$$d_n = \int_{-\frac{1}{2}}^{0} (2x + \frac{1}{2}) e^{-2i\pi nx} dx + \int_{0}^{\frac{1}{2}} (-2x + \frac{1}{2}) e^{-2i\pi nx} dx = \begin{cases} 0 \text{ si } n \text{ est pair} \\ \frac{2}{\pi^2 n^2} \text{ si } n \text{ est impair} \end{cases}$$

Ainsi les d_n décroissent en l'ordre de grandeur $1/n^2$ alors que les c_n décroissent comme $1/n$. Cela signifie que, pour approximer la fonction g (en norme 2) il faudra garder beaucoup moins de coefficients de Fourier que pour approximer la fonction f. Cette capacité à approximer plus efficacement une fonction symétrique fait que la DCT (qui symétrise avant de calculer la TFD) est plus efficace dans le contexte de la compression des signaux que l'on voit au chapitre suivant.

Chapitre 6

Compression des signaux naturels

Dans ce chapitre, nous nous intéressons à la compressibilité des signaux. Nous le ferons à travers des exemples concrets de signaux dits "naturels" c'est-à-dire destinés à être vus (images) ou écoutés (sons) par des êtres vivants.

Introduction

Nous appelons signaux naturels, les images et les sons. Ils sont destinés à véhiculer de l'information directement accessible à nos sens. Nous étudions dans ce chapitre les raisons qui font que ces signaux sont hautement compressibles et comment tirer parti de cette compressibilité pour proposer les bases des algorithmes de compression.

D'abord, voyons sur l'exemple d'une lettre de l'alphabet la grande redondance d'information véhiculée dans un signal naturel. Soit donc une lettre de l'alphabet. Pour coder cette lettre sur ordinateur on peut se contenter de noter son numéro dans l'alphabet (0 pour "A", 1 pour "B" et 4 pour "E"). Si le message ne contient que des caractères majuscules, on peut se contenter d'un codage sur 5 bits par caractère ($2^5 = 32 > 26$). Si nous voulons imprimer cette lettre pour qu'un humain la lise il faudra synthétiser une image (sur écran ou imprimante). Pour chaque caractère, il est généralement admis qu'il faut un minimum de d'une matrice 8×8 pixels pour imprimer de manière lisible. Si ces pixels ne sont que noirs ou blancs, il nous faut 64 bits pour chaque caractère. Si nous voulons en plus que le texte imprimé soit de belle facture, il faudra utiliser une matrice de 50×50 (police 12 points [1], en 300 dots per inch, standard minimal en bonne impression) avec au moins 16 dégradés de gris ce qui donne : $50.50.4 = 10000$ bits d'information sur papier pour chaque caractère. Nous avons multiplié par 2000 la quantité de bits entre le codage pour ordinateur et la version image d'un caractère.

Si nous disposons d'un logiciel capable, à partir d'une version numérisée d'un document, de retrouver, pour chaque caractère, sa police, sa taille, ce qu'il est (A, B...) alors on voit qu'une feuille A4 [2] scannée en 300 dpi, 16 niveaux de gris passe de $32Mbits \approx 3.10^7$ bits à $3000.5 = 15000$ bits (il y a à peu près 3000 caractères par page imprimée)

En un sens l'exemple ci-dessus résume tout ce chapitre, car nous chercherons à compresser les signaux naturels en les décomposants en atomes (les caractères dans l'exemple) qui se retrouvent souvent dans le signal et permettent de les exprimer efficacement.

1. point=1/72 inch
2. Le format A4 fait 1/16ème de m^2 et le pouce (inch) vaut 2,54 cm

Évidemment nous n'obtiendrons pas des taux de compression aussi forts que ceux envisagés ci-dessus, car les atomes que nous utiliserons par la suite ne sont pas aussi simples que les caractères et que les images sont souvent plus complexes qu'un document imprimé.

6.1 Définition (restreinte) de la compression

On ne s'intéresse qu'à la compression des signaux discrets et finis (il existe une théorie de l'approximation des signaux continus, mais elle est au-delà des objectifs de ce cours). Nous cherchons à exprimer un signal du mieux possible avec le moins d'atomes possible. Cela nous donne la définition suivante

Définition 6.1. Approximation, taux de compression
Soit x un vecteur de $\mathbb{R}^N$, $(\alpha_n)_{n\in\{1\dots M\}}$ une collection de M vecteurs de $\mathbb{R}^N$, $n_j \in \{1\dots M\}$ une collection d'indices pour j allant de 0 à $m-1$ et enfin les a_j (pour j dans le même intervalle) sont des coefficients réels. On note

$$\tilde{x} = \sum_{j=0}^{m-1} a_j \alpha_{n_j}$$

On dit alors que $\tilde{x}$ est une approximation de x dans la collection d'atomes α_n avec les coefficients a_j. Le **taux de compression** *τ_c est défini par*

$$\tau_c = \frac{m}{N}$$

*et l'***erreur relative de compression** *est définie par*

$$err_c = \frac{\|x-\tilde{x}\|}{\|x\|}$$

Dans l'espace $\mathbb{R}^N$ il est toujours possible de représenter un vecteur comme somme d'au plus N atomes. Il suffit pour cela d'écrire sa décomposition dans la base canonique. Cela explique notre définition du taux de compression. L'erreur de compression, quant à elle, reflète à quel point le signal est bien restitué par les m coefficients choisis.

Dans l'exemple introduction la famille α serait constituée des imagettes représentant tous les caractères de l'alphabet situés à toutes les positions possibles sur la feuille de papier et les coefficients a_l seraient soit 1 ou 0 suivant que le caractère est présent où pas à la position donnée.

Dans la suite, nous cherchons les meilleures familles α pour coder les signaux naturels. Nous cherchons aussi comment, une fois qu'une famille α est fixée, trouver pour un signal x donné, les coefficients a_j pour approximer x du mieux possible. Vous noterez que cela implique de faire une sélection parmi les atomes α entre ceux que l'on utilise et ceux que l'on n'utilise pas (au travers des indices n_j choisis).

Soit $x = (1,-1,1,-1)$ un vecteur de $\mathbb{R}^4$. Si on prends pour α_n la base canonique de $\mathbb{R}^4$ est que l'on prends $m = 2$, $n_0 = 1, n_1 = 4$ et $a_0 = 1/3, a_1 = -1$ on obtient

$$\tilde{x} = \frac{1}{3}(1,0,0,0) - 1.(0,0,0,1) = (\frac{1}{3},0,0,-1)$$

On alors un taux de compression de $1/2$ et une erreur relative de compression de $\frac{\sqrt{4/9+1+1+0}}{\sqrt{4}} \approx 0,78$

Une amélioration possible, sans augmenter le taux de compression, est de prendre $a_0 = 1$ ce qui donne

$$\tilde{x} = (1, 0, 0, -1)$$

et alors l'erreur relative de compression devient $\frac{\sqrt{0+1+1+0}}{\sqrt{4}} \approx 0,707$. On peut facilement montrer que pour cette famille α, ce vecteur et le taux de compression de $1/2$ on ne peut pas obtenir une plus faible erreur relative de compression (voir le cas des bases dans la suite).

Choisissons maintenant pour famille α, les vecteurs de Fourier que la TFD d'ordre 4. Ils ont la forme

$$\alpha_k = \left(1, e^{\frac{2i\pi k}{4}}, e^{\frac{4i\pi k}{4}}, e^{\frac{6i\pi k}{4}}\right)$$

pour k allant de 0 à 3. En particulier on remarque que pour $k = 2$ on a

$$\alpha_2 = (1, -1, 1, -1) = x$$

Ainsi, on a obtenu une écriture parfaite de x dans la base α qui ne fait intervenir qu'un seul terme non nul. Cela signifie que nous avons un taux de compression de $1/4$ avec une erreur relative de 0 (compression sans perte).

Cet exemple illustre l'importance du choix de la famille α.

6.2 Choix de la base α

Nous présentons une approche qui permet de construire une base α dans le but de minimiser le taux de compression et ce sur une base de données de signaux fixés qui servent d'exemples. Une fois la base construite à partir de cette base de données, on l'utilise pour la compression de signaux nouveaux qui ne font pas forcément partie de la base d'apprentissage.

Soit donc $V_1, \ldots, V_n$ une base de données de vecteurs de $\mathbb{R}^N$. On fixe m le nombre de coefficients non nuls et on appelle les $v_1, \ldots, v_m$ les vecteurs qui minimisent la quantité

$$E = \sum_{i=1}^{n} \|V_i - \sum_{j=1}^{m} \beta_j^i v_j\|^2$$

Les coefficients β_j^i sont libres et E représente le minimum d'erreur quadratique commise en approximant chaque V_i par une combinaison des v_i.

On remarque que pour i la quantité

$$\|V_i - \sum_{j=1}^{m} \beta_j^i v_j\|^2$$

est toujours plus grande (ou égale) à la distance (au carré) entre V_i et l'espace vectoriel engendré par les v_j. Elle est minimale si

$$\sum_{j=1}^{m} \beta_j^i v_j$$

est la projection orthogonale de V_i sur l'espace engendré par les v_j.

Ainsi, la valeur minimale de E ne dépend que de l'espace vectoriel engendré par les v_j. On peut donc supposer, sans pertes de généralité, que les v_i sont **orthonormés**. Dans ce cas, pour tout i, la projection orthogonale de V_i sur l'espace engendré par les v_j est

$$\sum_{j=1}^{m} < V_i | v_j > v_j$$

Et le problème devient : minimiser la quantité

$$E = \sum_{i=1}^{n} \|V_i - \sum_{j=1}^{m} < V_i | v_j > v_j\|^2$$

Si on note A la matrice

$$A = \sum V_i V_i^T$$

alors E s'écrit

$$E = \sum_i \|V_i\|^2 - \sum_{j=1}^{m} v_j^T A v_j$$

Or A est une matrice symétrique positive. On appelle $e_1, \ldots, e_N$ ses vecteurs propres (orthonormés) et $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \lambda_N$ les valeurs propres associées (on a ordonné les vecteurs propres pour qu'ils correspondent à des valeurs propres décroissantes).

Pour minimiser E, il faut maximiser

$$E' = \sum_{j=1}^{m} v_j^T A v_j$$

Si on fait l'hypothèse simplificatrice que les λ_i sont distincts deux à deux, alors la solution au problème de maximisation est :

L'espace vectoriel engendré par les $v_1, \ldots, v_m$ doit être le même que l'espace vectoriel engendré par les $e_1, \ldots, e_m$.

En effet, si on note $\gamma_i^j =< v_j | e_i >$ les coefficients de décomposition des v_j sur les e_i alors on a

$$E' = \sum_{j=1}^{m} \sum_i \lambda_i \left(\gamma_i^j\right)^2 = \sum_i \lambda_i \left(\sum_{j=1}^{m} (\gamma_i^j)^2 \right)$$

Or,

$$\sum_{j=1}^{m} (\gamma_i^j)^2$$

est la norme (au carré) de la projection de e_i sur l'espace engendré par les v_j et donc

$$\sum_{j=1}^{m} (\gamma_i^j)^2 \leqslant 1$$

par ailleurs,

$$\sum_i \sum_j (\gamma_i^j)^2 = m$$

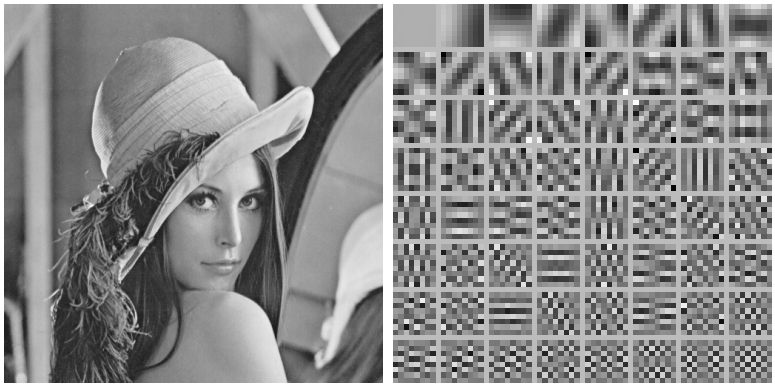


FIGURE 6.1 – À gauche une image dont on a extrait toutes les imagettes 8x8 que l'on a mises sous forme vecteur pour calculer la matrice A. À droite, nous avons placé les vecteurs propres de la matrice A dans l'ordre décroissant des valeurs propres qui leur sont associées.

(c'est la somme des carrés des normes des v_j)
Donc le maximum de E' est atteint lorsque

$$\forall 1 \leqslant i \leqslant m, \quad \sum_{j=1}^{m} (\gamma_i^j)^2 = 1$$

et vaut

$$\sum_{i=1}^{m} \lambda_i$$

6.2.1 Application

On se donne une image (figure 6.1). On extrait tous les morceaux de taille 8×8 et on leur applique ce que l'on vient de voir. On obtient ainsi la liste des $e_1, \ldots, e_N$ vecteurs propres de la matrice A. On classe ces vecteurs dans la figure de droite de gauche à droite de haut en bas dans l'ordre décroissant des valeurs propres associées.

Le graphique 6.2, montre, en fonction de m, la somme des $\lambda_1, \ldots, \lambda_m$. Cette somme est d'autant plus grande que les m premiers vecteurs permettent d'approximer bien tous les vecteurs V_i de la base de données.

Le fait qu'un seul vecteur capte 98% de l'énergie quadratique (première valeur du graphique) et la montée rapide du graphe vers 1 est une preuve du caractère parcimonieux des images (nous aurions les mêmes constatations pour du son). C'est-à-dire qu'il existe des bases bien adaptées pour lesquels les images et les sons font porter la plus grande partie de leur énergie sur peu de vecteurs de la base.

On remarque que la base obtenue ressemble à la base de Fourier. Cependant, elle ne lui est pas tout à fait équivalente. On peut remarquer, par exemple, que le vecteur e_{64} (en bas à droite de 6.1) est légèrement atténué sur ses bords. Cette différence est due au fait que les vecteurs V_i ne sont pas invariants par translation (circulaire). En effet, l'invariance par translation circulaire de l'ensemble des signaux V_i (i.e. toute translation (circulaire) de V_i est un autre vecteur V_j de la base) impliquerait que la matrice A est circulante (i.e.

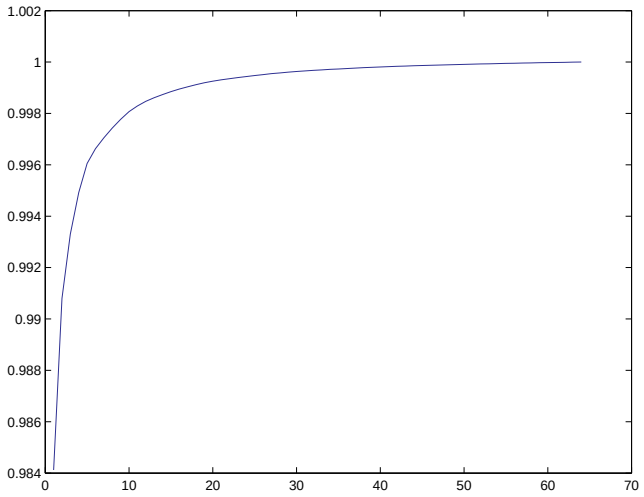


FIGURE 6.2 – Figure montrant la somme des $\lambda_1 + \lambda_2 ... \lambda_m$ en fonction de m. Cette somme est proportionnelle à la quantité d'énergie que captent les m premiers vecteurs de la figure 6.1. Remarquer que pour $m = 1$ on a déjà 98% de l'énergie représentée par un seul vecteur (le vecteur de fréquence 0).

est la matrice d'un SLI de $\mathbb{R}^N$ vers $\mathbb{R}^N$). Or, toute matrice circulante est diagonalisable dans la base de Fourier (car tout SLI est une convolution et qu'une convolution est un produit dans la base de Fourier [3]). Cependant, l'hypothèse d'invariance par translation circulaire n'est pas très réaliste pour des signaux naturels. Généralement la base de la DCT est préférée à calle de la TFD en traitement du signal.

6.3 Compression linéaire

La compression linéaire consiste à se donner une base orthonormée α_n ordonnée et à faire ceci :

Pour un taux de compression $\tau = \frac{m}{N}$ fixé et un signal $x \in \mathbb{R}^N$ on prends pour approximation de x le vecteur

$$\tilde{x} = \sum_{i=1}^{m} < x|\alpha_i > \alpha_i$$

Autrement dit, on choisit les m premiers vecteurs de la base α pour approximer x.

Cette compression est dite linéaire, car, une fois que le taux de compression est fixé, le vecteur $\tilde{x}$ dépend linéairement de x.

L'ordre des vecteurs α est important et on le détermine généralement grâce à une étude statistique sur une base de signaux. Par exemple on peut prendre

$$\alpha_i = e_i$$

3. La base de Fourier a des vecteurs à composante complexe, cependant, comme A est positive, les valeurs propres sont réelles et on en déduit facilement que les parties réelles et imaginaires des ondes de Fourier sont des vecteurs propres.

FIGURE 6.3 – Lena compressée en prenant les 5000 atomes de DCT (512x512) de plus basse fréquence.

où les e_i sont les vecteurs construits à la section précédente pris dans l'ordre décroissant des valeurs propres de A. Ceci garantit que les m premiers vecteurs engendrent l'espace vectoriel de dimension m qui approxime le mieux la base de données V_i. Cependant, étant donné un signal particulier x, rien ne garantit que l'utilisation des m premiers vecteurs $e_1, \ldots, e_m$ produit la meilleure approximation. C'est l'objet de la section suivante de proposer une compression qui adapte le choix des atomes au signal x qu'il faut compresser.

Exemple : Dans la figure 6.3, on a compressé l'image 6.1 en ne gardant que les 5000 coefficients de plus basse fréquence dans la base de la DCT2D 512x512 (la taille de l'image est 512x512). On remarque l'apparition de l'effet de Gibbs qui est caractéristique du filtrage passe-bas des images. Dans la figure **??**, nous avons utilisé une DCT 16x16 (ce qui 512*512/(16*16)=1024) et gardé 5 coefficients basse fréquence dans chaque bloc.

6.4 Compression adaptative sur une base

Dans cette section on suppose que la famille α forme une base de $\mathbb{R}^N$ (en particulier $M = N$) et on suppose de plus que cette famille est **orthonormée**. Contrairement à la section précédente, on va choisir les m vecteurs qui approximent le mieux x sans se contraindre à un ordre prédéfini sur les vecteur α_n.

On a le résultat suivant qui nous dit, dans le cas où α est une base orthonormée et pour un taux de compression fixé, comment choisir la meilleure approximation de x

Proposition 6.2. *Soit α_n une base orthonormée de $\mathbb{R}^N$ et x un vecteur de $\mathbb{R}^N$. Soit $m \leqslant N$ un entier fixé et soient les $c_n =< x|\alpha_n >$ les produits scalaires de x avec les vecteurs de la base α. On fixe le nombre de coefficients m de l'approximation. On note σ_n la permutation des indices $\{0 \ldots N-1\}$ telle que*

$$\forall n < N-1, \quad \left|c_{\sigma_{n+1}}\right| \leqslant |c_{\sigma_n}|$$

FIGURE 6.4 – Lena compressée en prenant les 5120 atomes de DCT 16x16 de plus basse fréquence(dans chaque bloc 16x16 on a gardé les 5 coefficients de plus basse fréquence). Les effets de Gibbs ont disparu, mais apparaît un effet de bloc.

c'est-à-dire que l'on a ordonné les produits scalaires, c_n par ordre décroissant de valeur absolue. Alors pour tout choix d'indices $n_0 \dots n_{m-1}$ et tout choix de coefficients $a_0, \dots, a_{m-1}$ on a

$$\|x - \sum_{j=0}^{j=m-1} a_j \alpha_{n_j}\| \geqslant \|x - x_m\|$$

avec

$$x_m = \sum_{j=0}^{j=m-1} c_{\sigma_j} \alpha_{\sigma_j}$$

Autrement dit, aucune compression de taille m de x n'est meilleure que la compression consistant à ne garder que les m plus grands produits scalaires de x contre les éléments de la base α.

Ce résultat, intuitif, nous dit que si la famille α est une base orthonormée et si le budget de compression m est fixé, la meilleure approximation de x est la somme des m plus grandes composantes de x dans la base α.

Démonstration. Comme la famille α est une base orthonormée, on sait que tout vecteur y vérifie

$$y = \sum_n < y|\alpha_n > \alpha_n$$

et

$$\|y\|^2 = \sum_n | < y|\alpha_n > |^2$$

Soit $J = \{n_0, \dots, n_{m-1}\}$ l'ensemble des indices pour lesquels les a_j sont non nuls. le cardinal de J est m.

On note $y = \sum_{j=0}^{j=m-1} a_j \alpha_{n_j}$ et on veut montrer que c'est une moins bonne approximation que x_m (défini comme la somme des m plus grandes composantes orthogonales de x). C'est-à-dire que l'on veut montrer que

$$\|y - x\|^2 \geqslant \|x - x_m\|^2$$

D'une part, par définition de x_m,

$$\|x - x_m\|^2 = \sum_{j>m} |c_{\sigma_j}|^2$$

(i.e. la somme les $N - m$ plus petits produits scalaires) et d'autre part

$$\|x - y\|^2 = \sum_n | < x - y|\alpha_n > |^2 = \sum_n |c_n - < y|\alpha_n > |^2 = \sum_{n \in J} |c_n - a_{n_j}|^2 + \sum_{n \notin J} |c_n|^2$$

Or, le dernier terme, $\sum_{n \notin J} |c_n|^2$ est une somme de $N - m$ termes de la forme $|c_n|^2$. Ce terme est donc plus grand que, $\sum_{j>m} |c_{\sigma_j}|^2$ car il s'agit de la somme des $N - m$ plus petits termes c_n. On a donc montré que

$$\|x - y\|^2 \geqslant \|x - x_m\|^2$$

□

6.4.1 Exemples

Dans la figure 6.5 nous présentons les résultats de la compression adaptative. Nous avons utilisé deux bases : la DCT 512x512 pour la figure de gauche et la DCT 16x16 pour la figure de droite. Les équivalents non adaptatifs (linéaires) sont les figures 6.4 et 6.3. Dans le cas DCT 512, on constate la disparition de l'effet de Gibbs et dans le cas DCT 16 on constate que les petits détails sont mieux reconstruits. Cela est dû au fait que nous laissons libre le choix des atomes à garder et que nous ne forçons pas le fait de garder que les atomes basse fréquence. Ainsi, les détails des plumes du chapeau vont avoir le droit de prendre un coefficient pour eu, et comme le budget m est constant, certains blocs 16x16 seront privés d'un des 5 coefficients que nous leur avions alloués pour la figure **??**. Par contre, on constate sur la figure de gauche l'apparition d'un "grain". Il est dû à la non localité des atomes 512x512.

6.5 Insuffisance des bases pour la capture efficace de l'information : Compromis entre localisation spatiale et fréquentielle

6.5.1 Localisation de bases pour des images

Jusqu'ici nous avons considéré la compression comme un processus de choix parmi une base des vecteurs qui portent le plus d'information et avons compressé un signal en annulant les coefficients faibles dans cette base. Cependant, si nous observons l'image de

FIGURE 6.5 – Lena compressée : gauche : en prenant les 5000 atomes de DCT 512x512 les plus forts. Droite : en prenant les 5000 atomes de la DCT 16x16. Ces figures sont à comparer à 6.4 et 6.3. On constate une nette amélioration visuelle, bien que le budget de coefficients non nuls n'ait pas changé.

FIGURE 6.6 – À gauche l'image "barbara". Elle contient des zones de textures localisées. à droite une compression sur une base trop localisée : la base des pixels (5000 pixels les plus lumineux).

gauche de la figure 6.6 , on se rend compte qu'elle contient des textures haute fréquence, mais que ces textures sont localisées en espace. Si nous utilisons la base de la DCT2D de taille 512x512, nous allons pouvoir capturer ces structures dans par les vecteurs haute fréquence. Cependant, les vecteurs haute fréquence de la DCT2D 512x512 ont un support qui s'étend sur l'image tout entière et vont être visibles sur toute l'image compressée $\tilde{x}$ même dans les zones plates de l'image.

Une solution à ce problème est de considérer une base de DCT locale d'une taille inférieure à 512x512, par exemple de taille 8x8.

Jusqu'où aller dans la localisation ? Si nous poussons le raisonnement de nécessité de la localisation jusqu'à son terme, nous arrivons à la plus fine localisation spatiale, c'est-à-dire une base dont les vecteurs sont les pixels. La figure de droite de 6.6 montre la compression la meilleure possible pour l'image de gauche sur la base des pixels. Cela revient à choisir les 5000 pixels les plus lumineux.

D'un autre côté, il se peut qu'un détail haute fréquence ait une taille plus grande que 8x8. Si nous fixons, par avance une taille à la DCT locale (ou n'importe quelle base localisée) il se pourrait que la compression soit moins efficace que pour une autre taille.

Il faudrait laisser le paramètre de taille de la base localisée libre. Pour cela nous aurons recours à plusieurs bases localisées de tailles de support différentes. Cependant l'union de la base DCT 8x8 et de la base DCT 16x16 n'est pas une base et le choix d'une décomposition de l'image dans l'union des deux n'est plus unique. De ce fait l'algorithme de compression adaptative développé plus haut pour une base n'est plus utilisable.

6.5.2 Le plan temps fréquence pour les sons

Nous avons vu en cours et en TP que la transformée de Fourier à Court Terme était un outil puissant d'analyse des sons. Rappelons le contexte : Un signal u étant donné, on définit sa TFCT par la formule

$$\forall (n,k) \in \mathbb{Z} \times \{0,\dots,M-1\},\ U\left(n, \frac{k}{M}\right) = \sum_{m\in\mathbb{Z}} u_m w_{m-n} e^{-2i\pi \frac{k}{M} m}$$

Dit autrement, $U(n, k/M)$ est le produit scalaire du signal u contre une troncature d'une onde de Fourier de fréquence k/M située autour de la position n. Le graphique 6.7 montre une telle TFCT.

D'un autre coté le graphique 6.8, montre le TFCT d'une onde de Fourier (non tronquée). Ici encore, on perçoit qu'une onde de Fourier (ou la base de la TFD de même taille que le signal complet) pourra capturer des raies de 6.7, mais son support temporel étant trop grand par rapport aux raies d'énergie de 6.7, elle ne les capture pas efficacement. Elle induira une onde parasite sur tout le reste du domaine temporel qu'il faudra essayer d'annuler avec d'autres atomes.

Par ailleurs, on sait que la localisation temporelle et fréquentielle sont antinomique : aucune fonction ne peut être à la fois à support très restreint en temps et en fréquence (principe d'incertitude de Heisenberg).

Le problème de la compression des signaux peut se voir comme une optimisation de support temps-fréquence des atomes utilisés pour approximer un signal.

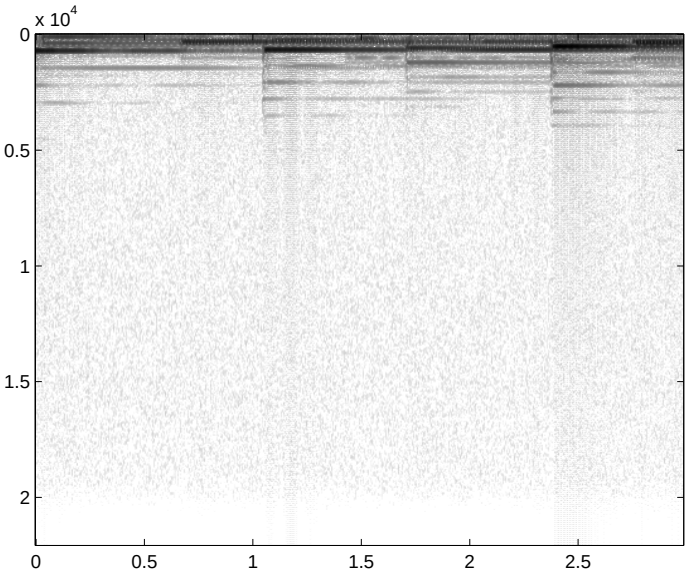


FIGURE 6.7 – **Les raies horizontales représentent des ondes d'un certain support temporel. Elles peuvent apparaître de manière arbitraire dans le temps (quand une note est jouée) et avoir des longueurs variables.**

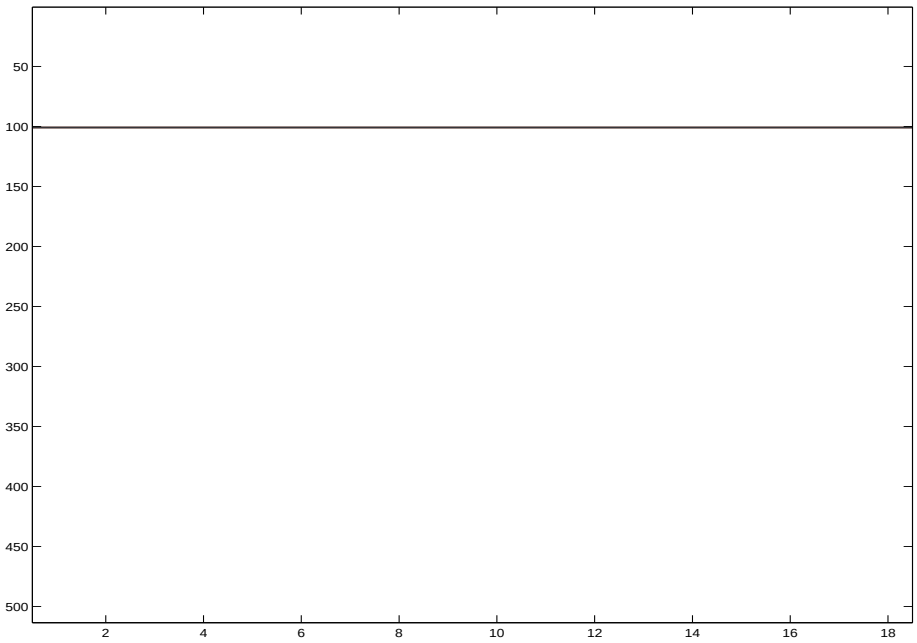


FIGURE 6.8 – **Spectrogramme d'une onde de support la taille du signal. Elle ne peut pas capturer de manière efficace une raie du spectrogramme tel que celui de la figure 6.7**

6.6 Cas des dictionnaires et algorithme des appariements successifs (Matching Pursuit)

La dernière méthode de compression que nous présentons permet l'utilisation d'un ensemble d'atomes α_n qui n'est pas forcément une base. Nous supposons quand même que les atomes α_n sont tous de norme 1 (changer la norme de α_n peut être compensé par le changement du coefficient qui lui est affecté pour produire l'approximation $\tilde{x}$, on ne perd donc rien à faire cette hypothèse).

Si les α_n ne forment plus une base, la question de trouver le meilleur approximant à m coefficients non nuls n'est plus triviale. Il ne suffit plus de prendre les plus grands $< x|\alpha_n >$. Pour cela un algorithme appelé Matching Pursuit (ou approximations successives) a été proposé par Stéphane Mallat et qui procède de la manière suivante

1. Poser $x_0 = x$ et $\tilde{x}_0 = 0$. On se donne un budget de coefficients m (taux de compression).
2. À l'étape j, calculer tous les produits scalaires $< x_j|\alpha_n >$.
3. Trouver n qui maximise $< x_j|\alpha_n >$. On l'appelle n_j.
4. Poser $x_{j+1} = x_j - < x_j|\alpha_{n_j} > \alpha_{n_j}$ et $\tilde{x}_{j+1} = \tilde{x}_j + < x_j|\alpha_{n_j} > \alpha_{n_j}$.
5. Retour à l'étape 2 en incrémentant j jusqu'à avoir $j = m$.
6. L'approximation $\tilde{x}_m$ est le résultat de l'algorithme.

Autrement dit, à chaque étape on cherche l'atome α_n qui capture le plus d'énergie du signal x_j. On retranche à x_j sa projection sur cet atome et on ajoute à $\tilde{x}_j$ cette même projection.

Proposition 6.3. *À chaque étape j de l'algorithme on a*

$$\tilde{x}_j + x_j = x_0 = x.$$

De plus, la norme de x_j décroit en fonction de j.

Enfin l'erreur d'approximation est

$$x - \tilde{x}_m = x_m$$

Démonstration. La première propriété est évidente par récurrence.

Le fait que

$$\|x_{j+1}\| \leqslant \|x_j\|$$

est du au fait que $\|y - < y|\alpha > \alpha\| \leqslant \|y\|$ est vrai pour tout vecteur y et vecteur α de norme 1. (il s'agit de retrancher à y sa projection orthogonale sur α, ce qui ne peut que faire diminuer la norme par le théorème de Pythagore) □

6.6.1 Exemple

Pour définir un algorithme de matching pursuit il suffit de définir le dictionnaire des atomes α que l'on utilise.

Pour l'expérience présentée à la figure 6.9, nous avons concaténé les bases DCT 64, 32, 16 et 8 et appliqué l'algorithme de matching pursuit avec 5000 pas ($m = 5000$).

FIGURE 6.9 – Gauche : image barbara compressée de manière adaptative sur la base DCT 8x8. droite : Résultat de l'algorithme matching pursuit avec le même budget de coefficients. On constate que l'algorithme matching pursuit a su bien reconstruire plus de détails haute fréquence. En effet, pour la DCT 8x8, si un détail haute fréquence est plus grand (en support) que 8x8, il faudra plusieurs coefficients pour le reconstruire alors que dans le dictionnaire utilisé pour le matching pursuit, il existe des atomes de tailles plus grande qui peuvent être utilisés.

Nous mettons le résultat en comparaison de l'algorithme de compression adaptative sur la DCT 8x8 avec le même taux de compression. Là où la DCT 8x8, n'a pas eu assez de "budget" pour coder la partie droite de la nappe, l'algorithme du matching pursuit, en adaptant la taille des atomes peut allouer des atomes de grand support lorsqu'un détail haute fréquence est de taille plus grande que 8x8.

6.6.2 Considérations algorithmiques

Le plus grand défaut de l'algorithme Matching Pursuit est le temps de calcul.

Si le signal est de taille N et que le dictionnaire est la concaténation de l DCT locales (de tailles différentes). Il faut , a priori, pour chaque étape recalculer toutes les transformées. Cela fait un temps de calcul de

$$mlN\log(N)$$

Cependant, on peut gagner du temps de calcul en ne recalculant pas tous les coefficients des décompositions après la mise à jour de x_j.

En effet, si l'atome choisi à l'étape j est α_{n_j}, il se peut que pour un grand nombre d'indices n le produit scalaire

$$< \alpha_{n_j} | \alpha_n >$$

soit nul. Et dans ce cas, le coefficient de $< x_{j+1} | \alpha_n >$ sera égal à $< x_j | \alpha_n >$, car x_j et x_{j+1} ne diffèrent que α_{n_j}.

Dans la pratique : Si le dictionnaire utilisé est une concaténation de vecteurs de supports variables, très peu de vecteurs du dictionnaires auront des supports d'intersection non vide. Si on choisit un atome DCT 8x8 à l'étape j, alors tous les blocs 8x8 où ne se situent pas l'atome ne seront pas affectés.

6.7 Liens avec les standards de compression

Dans ce chapitre nous avons vu comment utiliser le fait que les signaux naturels peuvent se décomposer de manière parcimonieuse sur certains vecteurs bien choisis pour réduire la place qu'il faut pour les représenter (sans perdre trop de détails).

De nombreux standards de compression existent, ils ont tous comme base de fonctionnement la transformation dans un plan temps-fréquence (ou espace-fréquence) du signal naturel, et le fait que les signaux y sont représentés de manière parcimonieuse.

Nous n'avons pas vu comment faire une étape de compression, celle qui consiste à quantifier les coefficients de $\tilde{x}$ et à les coder de manière efficace. Les coefficients de $\tilde{x}$ sont, ici, des réels or sur support informatique ces réels seront quantifiés. Par ailleurs, lorsqu'un atome α_n est choisi pour appartenir à la représentation $\tilde{x}$, nous n'avons pas dit comment coder son numéro n. Or, ce codage peut-être très consommateur de place mémoire (autant que le coefficient devant α_n). C'est essentiellement ces deux problématiques qui manquent pour produire un standard complet de compression.

Dans cette section, nous présentons le lien entre les standards de compression et ce que nous avons vu dans ce chapitre.

6.7.1 Codage JPEG pour les images

C'est de loin le standard de compression le plus répandu. Il est, hormis la quantification et le codage, équivalent à une compression adaptative sur la base de DCT2D 8x8. Le fait d'avoir fixé la taille de la DCT2D, bien qu'il fasse perdre un peu d'efficacité de compression permet l'implémentation très rapide de la compression.

6.7.2 Codage JPEG2000 pour les images

Dans le standard JPEG2000, la base utilisée est une base d'ondelettes. Les ondelettes sont une base obtenue en zoomant deux vecteurs de départ, l'un appelé passe-bas et l'autre passe-haut. Par exemple, en monodimenionnelle on part des deux vecteurs de $\mathbb{R}^2$: (-1,1) et (1,1). Ils forment bien une base de $\mathbb{R}^2$. On construit la base de $\mathbb{R}^4$ suivante : (1,1,1,1), (-1, -1 , 1 ,1) et (-1,1, 0, 0) et (0, 0, -1, 1). En d'autres termes pour passer la base d'ondelettes de $\mathbb{R}^{2^n}$ à la base d'ondelettes de $\mathbb{R}^{2^{n+1}}$ on retranche des deux bases le vecteur basse fréquence (1,...,1) , on réalise l'union des deux bases (l'une décalée de 2^n) cela nous donne $2(2^n - 1) = 2^{n+1} - 2$ vecteurs. On y ajoute le vecteur constant et le vecteur (1,1...,1,-1,-1,...-1) pour obtenir 2^{n+1} vecteurs.

La construction d'une base d'ondelettes revient à partitionner l'espace temps-fréquence de manière à ce que les vecteurs de grand support temporel (ou spatiale) codent les basses fréquences alors les vecteur de petit support codent les hautes fréquences.

Elle permet par ailleurs d'avoir accès, en une seule décomposition, à des atomes de différentes tailles.

Pour le reste, le codage JPEG2000 est similaire au codage JPEG.

6.7.3 Codage mp3 (et apparentés : ogg, acc) pour le son

C'est le codage son le plus répandu.
Pour le résumer nous dirons ceci :

1. On dispose de résultats de psychoacoustique qui disent que telle fréquence entendue avec telle énergie relative cache telle fréquence avec telle énergie. Cela donne une table de masquage : Pour chaque fréquence et chaque puissance, on connaît la liste des fréquences qui ne seront pas entendues si elle sont jouées simultanément.
2. On décompose localement un signal sur une base de Fourier locale. On consulte la table de masquage pour décider d'éliminer les fréquences masquées. Les fréquences masquantes sont elles incorporées au signal $\tilde{x}$.
3. On réitère en respectant des contraintes de recouvrement entre les fenêtres d'analyse.

Cet algorithme ressemble à celui du Matching pursuit avec la différence que l'on efface de x non seulement l'atome de plus forte énergie, mais également les atomes qui sont masqués par les atomes choisis.

Chapitre 7

Processus aléatoires sur $\mathbb{Z}$

7.1 Introduction

Dans ce chapitre, nous introduisons les processus aléatoires. Il s'agit de la modélisation mathématique de phénomènes imprédictibles. On utilise la notion de variable aléatoire pour arriver à une définition de signal aléatoire. Une variable représentera une des valeurs du signal et un processus sera une collection de variables aléatoires indexées par le temps (temps discret dans ce chapitre). On étudiera en détail une classe de processus dits SSL qui ont la propriété d'être invariants par translation et pour lesquels les outils de la théorie de Fourier et de la convolution vont pouvoir s'appliquer grâce à cette invariance dans le temps.

7.2 Définition des processus

Vous avez vu en cours de probabilité ce qu'est une variable aléatoire. La notion de processus s'intéresse aux rapports entre toute une famille de variables aléatoires. Dans ce cours on se limite à une famille de variables aléatoires indexée par $\mathbb{Z}$. Ce sont les processus discrets, que nous appelons simplement processus ici. Il existe une théorie des processus définis sur $\mathbb{R}$ mais elle est trop compliquée pour être présentée dans ce cours. Néanmoins, les outils définis ici peuvent servir d'intuition pour manipuler des processus continus. Notion dont vous vous servirez dans le cours de communication numérique.

On se donne dans la suite un espace probabilisé Ω sur lequel est définie une mesure de probabilité $\mathbf{P}$. On rappelle que

- $\mathbf{P}(\Omega) = 1$. Ω est de mesure fini.
- Une variable aléatoire est une fonction mesurable de Ω vers $\mathbb{C}$ (ou $\mathbb{R}$, variable réelle).
- Si X est une variable aléatoire. On dit que X possède une moyenne si

$$\int_\Omega |X(\omega)| d\mathbf{P}(\omega) < +\infty$$

on dit que $X \in L^1(\Omega)$ et on définit sa moyenne par

$$E(X) = \int_\Omega X(\omega) d\mathbf{P}(\omega)$$

— Si X possède une moyenne, on appelle X *centrée* et on note X^c la variable X à laquelle on a retranché sa moyenne

$$\forall \omega \in \Omega,\ X^c(\omega) = X(\omega) - E(X)$$

X^c possède une moyenne nulle : $E(X^c) = 0$. Ici on réserve la notation $\overline{X}$ à la conjugaison des complexes, alors que cette notation est parfois utilisée pour noter la moyenne d'une variable aléatoire. La moyenne est aussi appelée espérance.

— Si X est une variable aléatoire. On dit que X possède une variance si $X \in L^2(\Omega)$ c'est-à-dire

$$\int_\Omega |X(\omega)|^2 d\mathbf{P}(\omega) < +\infty$$

Dans ce cas on sait que $X \in L^1(\Omega)$ (car Ω est de mesure finie, comme on a déjà vu $L^2([-\frac{1}{2}, \frac{1}{2}[) \subset L^1([-\frac{1}{2}, \frac{1}{2}[)$). On définit sa variance par

$$Var(X) = \int_\Omega |X(\omega) - E(X)|^2 d\mathbf{P}(\omega)$$

Définition 7.1. Les processus
Un processus X est une fonction de $\mathbb{Z}$ vers l'ensemble des variables aléatoires. C'est une suite de variables aléatoires. C'est-à-dire que pour tout $n \in \mathbb{Z}$, X_n est une variable aléatoire.

Définition 7.2. Les processus stationnaires à l'ordre 1
Si X est un processus tel que $\forall n, X_n \in L^1(\Omega)$. On dit que X est stationnaire à l'ordre 1 si

$$\exists m_X \in \mathbb{C},\ \forall n \in \mathbb{Z},\ E(X_n) = m_X.$$

Autrement dit, la moyenne de toutes les variables X_n est la même. Nous notons cette moyenne m_X

Définition 7.3. Covariance de deux variables
Si X et Y sont deux variables qui possèdent des variances (elles sont L^2) on appelle covariance de X et Y et on note $Cov(X, Y)$ la grandeur

$$Cov(X, Y) = E\left[(X - E(x))\overline{(Y - E(Y))}\right] = E(X^c\overline{Y^c})$$

D'après l'inégalité de Cauchy-Schwartz, on a

$$|Cov(X, Y)| \leqslant \sqrt{Var(X)Var(Y)}$$

Définition 7.4. Processus stationnaire à l'ordre 2
On dit que le processus X est stationnaire à l'ordre 2 si $\forall n \in \mathbb{Z},\ X_n \in L^2(\Omega)$ et que l'on a

$$\forall k \in \mathbb{Z}, \forall n \in \mathbb{Z},\ Cov(X_{n+k}, X_n) = Cov(X_k, X_0)$$

Autrement dit, la covariance entre deux variables de ce processus ne dépend que de la distance entre les deux (c'est-à-dire de k) et non pas de leur position absolue (c'est-à-dire n). D'où le terme de stationnarité.

En particulier, la variance de toutes les variables X_n est la même (c'est la propriété ci-dessus avec $k = 0$).

Définition 7.5. Processus stationnaire au sens large (SSL)
Les processus stationnaires au sens large (SSL) sont les processus stationnaires à l'ordre 1 et 2. En particulier, il faut que les variables X_n aient une variance finie.

Définition 7.6. Autocovariance d'un processus stationnaire au second ordre
Si X est un processus stationnaire au second ordre, on appelle autocovariance de X et on note R_X la fonction définie sur $\mathbb{Z}$ (c'est une suite, mais on la note de manière fonctionnelle, car la place de l'indice est prise par X...) par

$$\forall k,\ R_X(k) = Cov(X_k, X_0) = Cov(X_{n+k}, X_n) = E((X_k - E(X_k))\overline{(X_0 - E(X_0))})$$

On a la propriété suivante, immédiate à démontrer : R_X possède la symétrie hermitienne,

$$\forall k \in \mathbb{Z},\ R_X(-k) = \overline{R_X(k)}$$

En particulier $R_X(0) \in \mathbb{R}$ est positive. Par ailleurs, R_X est maximale en 0

$$\forall k,\ |R_X(k)| \leqslant R_X(0)$$

Cela est du à l'inégalité de Cauchy-Schwartz.

Définition 7.7. Densité spectrale de puissance (DSP)
Si X est un processus stationnaire au second ordre et que R_X est son autocovariance. Si R_X est sommable, c'est-à-dire

$$\sum_{k\in\mathbb{Z}} |R_X(k)| < \infty$$

($R_X \in l^1$) on définit la densité spectrale de puissance de X, que l'on note S_X par

$$\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[,\ S_X(\nu) = \sum_{k\in\mathbb{Z}} R_X(k) e^{-2i\pi\nu k}$$

Autrement dit, S_X est la transformée de Fourier à temps discret (TFtD) de R_X.

En raison de la symétrie hermitienne du signal R_X, S_X est une fonction à valeurs réelles

$$\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[,\ S_X(\nu) \in \mathbb{R}$$

Définition 7.8. Puissance d'un processus SSL
Si X est un processus SSL on appelle sa puissance la norme L^2 au carré de X_n. On la note P_X. Elle ne dépend pas de n. Elle est donnée par la formule

$$P_X = E(|X|^2) = |m_X|^2 + R_X(0) = m_X^2 + \int_{-\frac{1}{2}}^{\frac{1}{2}} S_X(\nu) d\nu$$

La dernière égalité (qui suppose que R_X est sommable) est une application du théorème d'inversion de la TFtD.

Remarque 7.9. On ne parle pas d'énergie d'un processus, mais de puissance. En effet, pour tout processus SSL non nul, on peut montrer que pour presque tout ω la suite $n \mapsto X_n(\omega)$ n'est pas d'énergie finie. Par contre, la moyenne de $|X_n(\omega)|^2$ sera bien finie.

Remarque 7.10. Comme dans le cas de la transformation de Fourier, il est possible de généraliser la notion de DSP dans le cas où R_X est seulement l^2 ou encore aller plus loin et introduire la notion de distributions. Nous ne le faisons pas dans ce cours.

Remarque 7.11. Le nom "densité spectrale de puissance" suggère que S_X, en plus d'être à valeurs réelles est une fonction positive. C'est l'objet de la proposition suivante. Dans la démonstration qui suivra, nous utiliserons la relation suivante

$$R_X(k) = E(X_k^c \overline{X_0^c}) = \frac{1}{2}\left(E(X_k^c \overline{X_0^c}) + E(X_{k+1}^c \overline{X_1^c})\right) =$$

$$\frac{1}{N}\left(E(X_k^c \overline{X_0^c}) + E(X_{k+1}^c \overline{X_1^c}) + \cdots + E(X_{k+N-1}^c \overline{X_{N-1}^c})\right) =$$

$$E\left(\frac{1}{N}\sum_{t=0}^{t=N-1} X_{k+t}^c \overline{X_t^c}\right)$$

Les égalités successives étant obtenues par stationnarité à l'ordre 2. Cette égalité suggère (mais ce n'est qu'une intuition) que R_X est la convolution du processus X^c avec le processus $Y_n^c = \overline{X_{-n}^c}$. Or la transformée de Fourier du signal symétrique-conjugué est conjuguée de la transformée de Fourier du signal. La convolution étant transformée en produit par la TFtD, la transformée de Fourier de R_X sera le module au carré d'une transformée de Fourier et donc positive.

Proposition 7.12. Positivité de la DSP
Si X est un processus stationnaire au second ordre et que R_X est sommable alors

$$\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[,\ S_X(\nu) \geqslant 0$$

Démonstration.
Pour utiliser l'intuition introduite dans la remarque ci-dessus, nous allons considérer, pour $N \in \mathbb{N}$ le signal R_X^N définit comme la multiplication de R_X par une suite triangle

$$R_X^N(k) = \begin{cases} \frac{N-|k|}{N} R_X(k) \text{ si } |k| < N \\ 0 \text{ sinon} \end{cases}$$

On appelle S_X^N la TFtD de R_X^N. Le théorème de convergence dominée nous dit que S_X^N tend uniformément vers, S_X car R_X^N tend en norme l^1 vers R_X. Ainsi, si on prouve que S_X^N est positive sur $[-\frac{1}{2}, \frac{1}{2}[$ alors on aura montré que S_X l'est aussi. Dans ce qui suit nous montrons que S_X^N est positive

$$S_X^N(\nu) = \frac{1}{N}\sum_{k=-N+1}^{k=N-1} (N - |k|) R_X(k) e^{-2i\pi\nu k} =$$

$$\frac{1}{N}\sum_{k=-N+1}^{k=N-1} \left(\sum_{m=\max(0,-k)}^{m=\min(N-k-1,N-1)} E(X_{m+k}^c \overline{X_m^c})\right) e^{-2i\pi\nu(m+k-m)}$$

Le remplacement de $(N-|k|)R_X(k)$ par la somme sur m vient de la stationnarité de X. Par linéarité de l'espérance (E) on a que

$$S_X^N(\nu) = \frac{1}{N} E\left(\sum_{k=-N+1}^{k=N-1} \sum_{m=\max(0,-k)}^{m=\min(N-k-1,N-1)} X_{m+k}^c e^{-2i\pi(m+k)} \overline{X_m^c e^{-2i\pi\nu m}} \right)$$

Si on fait le changement de variable $n = m+k$ et $t = m$, on peut se convaincre que le couple (n,t) parcourt tout $\{0,\dots,N-1\}^2$ exactement une seule fois. En effet, la somme intérieure est un regroupement de tous les couples (n,t) qui vérifient $n-t=k$ et k parcourt $-N+1,\dots,N-1$ c'est-à-dire toutes les différences possibles entre n et t. On a donc

$$S_X^N(\nu) = \frac{1}{N} E\left(\left(\sum_{n=0}^{N-1} X_n^c e^{-2i\pi\nu n} \right) \overline{\left(\sum_{t=0}^{N-1} X_t^c e^{-2i\pi\nu t} \right)} \right) = \frac{1}{N} E\left(\left| \sum_{n=0}^{N-1} X_n^c e^{-2i\pi\nu n} \right|^2 \right) \geqslant 0$$

En effet, l'espérance d'une variable positive est positive. Au passage, la dernière équation est bien conforme à notre intuition que S_X est le produit de la TFtD de X par celle du processus symétrique conjugué de X. □

Exemple 7.13. **Le bruit blanc**
Si les X_n sont une suite de variables réelles indépendantes et identiquement distribuées de $L^2(\Omega)$, on dit que le processus X est un **bruit blanc**. On a, car les variables sont identiquement distribuées,

$$\forall n \in \mathbb{Z},\ E(X_n) = \int_{\mathbb{R}} t p_{X_n}(t) dt = \int_{\mathbb{R}} t p_{X_0}(t) dt = E(X_0)$$

où p_{X_n} est la densité de probabilité de la variable X_n (p_{X_n} ne dépend pas de n par hypothèse d'identique distribution). Le processus X est donc stationnaire à l'ordre 1.

de plus, si $k = 0$

$$E(X_{n+k}^c X_n^c) = E((X_n^c)^2) = \int_{\mathbb{R}} (t - E(X_n))^2 p_{X_n}(t) dt = \int_{\mathbb{R}} (t - E(X_0))^2 p_{X_0}(t) dt = E((X_0^c)^2)$$

Par égalité des densités de probabilité $p_{X_0} = p_{X_n}$.

Si $k \neq 0$

$$E(X_{n+k}^c X_n^c) = E(X_{n+k}^c) E(X_n^c) = 0$$

L'espérance d'un produit est le produit des espérances lorsque deux variables sont indépendantes. Le processus X est donc stationnaire au second ordre. Si on note $\sigma^2 = Var(X_0)$ son autocovariance est

$$R_X(k) = \sigma^2 \delta_k^0$$

et sa DSP est

$$\forall \nu [-\frac{1}{2}, \frac{1}{2}[,\ S_X(\nu) = \sigma^2$$

Ainsi, le bruit blanc porte de la puissance de manière équivalente pour toutes les fréquences ν .

Exemple 7.14. **un bruit blanc filtré**
On garde le processus bruit blanc définit ci-dessus et on note $Y_n = \frac{1}{2}(X_n + X_{n-1})$. Il est clair que $Y_n \in L^2(\Omega)$ (comme somme de deux variables L^2). On voit aussi qu'il est stationnaire à l'ordre 1 et à l'ordre 2. En effet

$$\forall n \in \mathbb{Z},\ E(Y_n) = \frac{1}{2}(E(X_n) + E(X_{n-1})) = E(X_0) = m_X$$

et

$$\forall k \in \mathbb{Z}, n \in \mathbb{Z},\ E(Y^c_{n+k}Y^c_n) = \frac{1}{2}R_X(k) + \frac{1}{4}R_X(k+1) + \frac{1}{4}R_X(k-1)$$

Qui ne dépend pas de n.

$$R_Y(k) = \begin{cases} \frac{1}{4} \text{ si } |k| = 1 \\ \frac{1}{2} \text{ si } k = 0 \\ 0 \text{ sinon} \end{cases}$$

Le processus Y peut être vu comme le filtrage du processus X par un filtre passe-bas (de réponse impulsionnelle $\ldots, 0, \frac{1}{2}, \frac{1}{2}, 0, \ldots$). Et, en effet, la DSP de Y (tracez-là) est bien plus forte dans les basses fréquences que dans les hautes. Le rapport entre la DSP d'un processus résultant d'un filtrage et la DSP du processus d'origine est vu dans la suite.

7.3 Filtrage des processus SSL

Proposition 7.15. filtrage par un filtre sommable
Soit X un processus SSL tel que R_X soit sommable et h une suite sommable. On appelle Y le processus filtré de X par le noyau h ce que l'on note $Y = h * X$. Le processus est défini par

$$\forall n \in \mathbb{Z},\ Y_n = \sum_{l \in \mathbb{Z}} h_l X_{n-l}$$

La somme ci-dessus étant prise dans $L^2(\Omega)$. Alors on a

1. *Pour presque tout $\omega \in \Omega$*

$$\forall n \in \mathbb{Z},\ Y_n(\omega) = \sum_{l \in \mathbb{Z}} h_l X_{n-l}(\omega)$$

   (ce qui signifie que pour presque tout tirage ω la somme $h_l X_{n-l}(\omega)$ a un sens, ce qui n'est a priori pas clair)
2. *Le processus Y est SSL. On note $\tilde{h}_n = \overline{h_{-n}}$ le signal h symétrisé et conjugué. On a*

$$m_Y = m_X \sum_{l \in \mathbb{Z}} h_l$$

$$R_Y = (h * \tilde{h}) * R_X$$

   Soit ponctuellement

$$\forall k \in \mathbb{Z},\ R_Y(k) = \sum_l (h * \tilde{h})(l) R_X(k-l) = \sum_{t,m} h_t \overline{h_m} R_X(k - t + m)$$

   et

$$\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[,\ S_Y(\nu) = |\hat{h}(\nu)|^2 S_X(\nu)$$

   ($\hat{h}$ est la TFtD de h)

3. *Si g est un autre signal sommable et que Z est le processus résultant du filtrage de du processus Y par g alors le processus Z est le résultat du filtrage de X par le signal sommable $h * g$. Ceci s'écrit*

$$(g * (h * X)) = (g * h) * X$$

Démonstration.
Pour tout l la fonction $\omega \mapsto h_l X_{n-l}(\omega)$ est bien dans $L^2(\Omega)$ et a pour norme $|h_l| \|X_{n-l}\|_2 = |h_l|\sqrt{P_X}$ où P_X est la puissance du processus X. Donc la suite de fonctions
$l \in \mathbb{Z} \mapsto (\omega \mapsto h_l X n - l(\omega))$ est sommable dans l'espace de Banach $L^2(\Omega)$. La définition de Y_n a donc bien un sens.

1. Le même raisonnement que ci-dessus permet de dire que $Y_n \in L^1(\Omega)$ est bien définie (comme somme des $h_l X_{n-l}$). D'abord, par Cauchy-Schawrtz on a

$$\int_\Omega |X_{n-l}(\omega)| d\mathbf{P}(\omega) \leqslant \sqrt{\int |X_{n-l}(\omega)|^2 d\mathbf{P}(\omega) \int_\Omega 1 d\mathbf{P}(\omega)} = \sqrt{P_X}$$

   On va appliquer le théorème de Fubini, calculons

$$\int_\Omega \left(\sum_{l \in \mathbb{Z}} |h_l| |X_{n-l}(\omega)| \right) d\mathbf{P}(\omega) = \sum_{l \in \mathbb{Z}} |h_l| E(|X_{n-l}|) \leqslant \|h\|_1 \sqrt{P_X} < \infty$$

   Ceci justifie le fait que pour presque tout ω

$$\sum_l h_l X_{n-l}(\omega)$$

   existe. Autrement dit que la série de fonctions $h_l X_{n-l}(\omega)$ converge presque partout vers une certaine fonction. Par unicité de la limite, cette limite ponctuelle ne peut être autre que Y_n.

   Ce résultat est a priori contre-intuitif. Pensez par exemple à un processus bruit blanc gaussien. On peut facilement prouver que la probabilité pour que la suite $X_n(\omega)$ soit bornée est nulle (une infinité de variables gaussienne i.i.d ne peuvent pas être bornées toutes en même temps par une même constante). Or, nous n'avons jamais envisagé de convoluer un signal sommable contre un autre signal qui ne serait même pas borné. La propriété que nous venons de démontrer nous dit que, bien que $X_n(\omega)$ ne soit pas bornée (lorsque n varie), la suite $h_l X_{n-l}(\omega)$ est néanmoins (presque) toujours sommable. On peut comprendre cela en se disant que les grandes valeurs de $X_n(\omega)$ ne sont que très rarement atteintes [1].

2. D'abord l'ordre 1

$$\forall n \in \mathbb{Z},\ E(Y_n) = E(\sum h_l X_{n-l}) = \sum_l h_l E(X_{n-l}) = \sum_l h_l m_X$$

   L'interversion entre l'espérance (qui est une intégrale sur Ω) et la somme sur l est justifiée par le théorème de Fubini. Nous ferons dans la suite ces interversions sans justification.

1. Penser à $g_n = n$ si n est une puissance de 2 et 0 sinon, et $h_n = 1/(n^2+1)$ le produit $h_n g_n$ est bien sommable, la somme est même inférieure à 2. Pourtant g n'est pas bornée

L'ordre 2 : D'abord $Y_n^c = Y_n - m_Y = \sum_l h_l(X_{n-l} - m_X) = \sum_l h_l X^c$. ainsi le processus Y centré est le filtrage du processus X^c par h. Dans la suite nous faisons l'hypothèse que X est centré (et donc Y aussi) pour alléger les notations.

$$E(Y_{n+k}\overline{Y_n}) = E(\sum_t h_t X_{n+k-t} \overline{\sum_m h_m X_{n-m}}) =$$

$$\sum_{t,m} h_t \overline{h_m} E(X(n+k-t)\overline{X(n-m)}) = \sum_{t,m} h_l \overline{h}_m R_X(k-t+m) =$$

(la dernière égalité en raison de la stationnarité de X)

$$\sum_{t,m} h_l \tilde{h}_{-m} R_X(k-(t-m)) = \sum_{t,m} h_l \tilde{h}_m R_X(k-(t+m)) =$$

$$\sum_r R_X(k-r) \left(\sum_{t+m=r} h_t \tilde{h}_m \right) = \left(R_X * (h * \tilde{h}) \right)(k)$$

Enfin la TFtD de $\tilde{h}$, vérifie $\hat{\tilde{h}}(\nu) = \overline{\hat{h}(\nu)}$ d'où la formule

$$S_Y(\nu) = |\hat{h}(\nu)|^2 S_X(\nu)$$

par les propriétés habituelles de la TFtD (S_Y est la TFtD de R_Y donc le produit des TFtD de h, $\tilde{h}$ et R_X $(= S_X)$

3.

$$Z_l(\omega) = \sum_m g_m Y_{l-m}(\omega) = \sum_m g_m \left(\sum_t h_t X_{l-m-t}(\omega) \right) = \sum_{m,t} g_m h_t X_{l-m-t}(\omega) =$$

$$\sum_k X_{l-k}(\omega) \left(\sum_{m+t=k} g_m h_t \right) = \sum_k X_{l-k}(\omega)(g*h)(k) = ((g*h)*X)_l(\omega)$$

par les propriétés précédentes.

□

Proposition 7.16. filtrage récursif
Si $b_0, \ldots, b_q$ et $a_0, \ldots, a_p$ sont des complexes tels que les polynômes

$$P(z) = \sum_n a_n z^n$$

et

$$Q(z) = \sum_n b_n z^n$$

n'ont pas de zéros communs et que Q n'a pas de zéros sur le cercle unité $\mathbb{U}$. Si de plus X est un processus SSL, alors il existe un unique processus SSL Y tel que

$$\sum b_i Y_{n-i} = \sum a_i X_{n-i}$$

et $Y = h * X$ où h est la réponse impulsionnelle du filtre récursif stable défini par cette équation de récurrence (voir chapitre sur la TZ).

Démonstration.
Existence : On appelle Z le processus

$$Z_n = \sum a_i X_{n-i}$$

Le processus Z comme convolution de X contre un filtre RIF (dont la réponse impulsionnelle est les a_i) est bien un SSL. Si on pose $Y = h * X$, alors Y est un SSL. D'après l'étude faite dans le chapitre sur la transformée en Z la convolution de h avec le signal fini b_i est égale au signal fini a_i. D'après la proposition précédente il vient que

$$\sum b_i Y_{n-i} = (b * Y) = (b * (h * X)) = (b * h) * X = a * X = Z$$

Donc Y est bien solution SSL de l'équation de récurrence.

Unicité : Soit Y un processus solution. Soit w la suite sommable telle que $w * b = \delta$. w existe bien, c'est la RI du filtre récursif dont le membre de droite est les b_i et le membre de gauche est δ (voir chapitre TZ). Alors, $w * (b * Y) = w * (a * X)$ (car $b * Y = a * X$) et donc

$$Y = (w * a) * X$$

donc Y est bien unique.

On rappelle que w a pour TZ $W(z) = \frac{1}{Q(z^{-1})}$ et h a pour TZ $H(z) = \frac{P(z^{-1})}{Q(^{-1})}$ et que dire que la convolution de deux signaux vaut δ signifie que le produit de leurs TZ est la fonction constante égale à 1. □

7.4 Applications

Munis des outils que nous venons de voir nous allons aborder des problèmes concrets pour les résoudre.

7.4.1 Prédiction linéaire pour le codage

On dispose d'une suite d'échantillons x_n et on cherche les coefficients $a_0, \ldots, a_p$ telles que

$$\epsilon = a * x$$

soit en norme quadratique la plus petite possible.

Si nous trouvons de tels coefficients alors on stockera sur le disque les coefficients $a_0, \ldots, a_p$ et la suite ϵ puis, pour décoder, on résoudra l'équation de récurrence (dont l'inconnue est la suite y)

$$a_0 y_n + a_1 y_{n-1} + \cdots + a_p y_{n-p} = \epsilon_n$$

Si les a_i sont bien choisis, y ne peut être que la suite x (voir chapitre sur la TZ).

D'où vient ce modèle ? Ce modèle simple vient d'une hypothèse sur la génération du son. Cette hypothèse est que le son, parole par exemple, est généré par les poumons qui envoient un bruit blanc dans la cavité résonnante (gorge, bouche, langue...) qui agit comme un SLI. Les coefficients a_i sont là pour modéliser ce SLI (ou plutôt son SLI inverse).

De cette explicitation du modèle, il vient qu'il faut régulièrement changer les coefficients a_i afin de refléter le fait que la cavité résonnante a changé de forme (passage à un autre phonème). Il faut donc découper un signal long (tout un discours) en morceaux d'une durée de l'ordre du centième de seconde.

Il est possible d'être encore plus générique et de trouver aussi des coefficients b_i qui filtrent ϵ (équation de récurrence complète), mais la présentation est plus compliquée.

Résolution

Nous adoptons ici l'approche par moindres carrés pour déterminer les coefficients a_i. On ne se demande pas comment choisir le p (nombre de coefficients non nuls) optimal.

On a donc une suite finie (découpage du signal) d'échantillons $x_0, \ldots, x_{N-1}$ d'échantillons (si le morceau fait 0.01s, cela donne de l'ordre de N=400 échantillons à la fréquence d'échantillonnage de 44100Hz). On note $\mathbf{x}$ le vecteur ligne dont les composantes sont les x_i (il est de taille N). On normalise les coefficients a_i par $a_0 = 1$ et on note $\mathbf{a}$ le vecteur ligne des a_i. Si on note ϵ le vecteur de $\epsilon_0, \ldots, \epsilon_{N-1}$ et $\mathbf{x}^k$ pour $k = 1 \ldots p$ les vecteurs ligne de taille N dont les premières composantes sont nulles et que les autres sont données par $\mathbf{x}_i^k = \mathbf{x}_{i-k} = x_{i-k}$. On a l'équation matricielle

$$\mathbf{a} \begin{bmatrix} \mathbf{x} \\ \mathbf{x}^1 \\ \cdots \\ \mathbf{x}^p \end{bmatrix} = \epsilon^T$$

Minimiser la norme de ϵ c'est minimiser la norme du vecteur

$$a_0\mathbf{x} + a_1\mathbf{x}^1 + \cdots + a_p\mathbf{x}^p$$

Nous avons forcé $a_0 = 1$ il nous reste donc à minimiser la quantité E_q définie par

$$E_q = \|\mathbf{x} + a_1\mathbf{x}^1 + \cdots + a_p\mathbf{x}^p\|_2^2 = (\mathbf{x} + a_1\mathbf{x}^1 + \cdots + a_p\mathbf{x}^p).(\mathbf{x} + a_1\mathbf{x}^1 + \cdots + a_p\mathbf{x}^p)^T$$

Le "." représente le produit matriciel, qui est ici réduit à un produit scalaire. Si on note M la matrice de taille $p+1 \times p+1$ dont les entrées sont $m_{i,j} < \mathbf{x}^i|\mathbf{x}^j >$ on a

$$E_q = \sum_{i=0}^{p}\sum_{j=0}^{p} a_i a_j m_{i,j} = \mathbf{a}M\mathbf{a}^T$$

Si on appelle L la sous matrice de M correspondant à $i > 1$ et $j > 1$ on a

$$E_q = \|\mathbf{x}\|_2^2 + 2\mathbf{a}_1\mathbf{V} + \mathbf{a_1}L\mathbf{a}_1^T$$

où $\mathbf{a}_1$ est le vecteur $\mathbf{a}$ privé de sa première composante $a_0 = 1$ (il est de taille p). et $\mathbf{V}$ est le vecteur colonne de taille p dont les composantes sont $< \mathbf{x}|\mathbf{x}^k >$ (pour $k > 0$).

Les coefficients $a_1, \ldots, a_p$ qui minimisent E_q sont données par

$$\mathbf{a}_1^T = -L^{-1}\mathbf{V}$$

Interprétation au regard des outils introduits

En multipliant L et V par le facteur $1/N$, l'équation ci-dessus peut se réécrire

$$\mathbf{a}_1^T = -(\frac{1}{N}L)^{-1}(\frac{1}{N}V) \tag{7.1}$$

Explicitons les coefficients du vecteur $1/N\mathbf{V}$, elles sont de la forme

$$V_k = \text{ pour } k = 1, \ldots, p, \ \frac{1}{N} < \mathbf{x}|\mathbf{x}^k >= \frac{1}{N}\sum_{l=k}^{N-1} x_l x_{l-k}$$

C'est une moyenne spatiale du produit $x_l x_{l-k}$. Si x est l'échantillonnage d'un processus centré réel, on a

$$E(V_k) = E(\frac{1}{N}\sum_{l=k}^{N-1} x_l x_{l-k}) = \frac{1}{N}(N-k)R_X(k)$$

Lorsque le nombre N devient grand, le vecteur V a pour espérance statistique un vecteur composé des $R_X(k)$.

De la même manière les coefficients de la matrice L ont pour espérance

$$E(L_{i,j}) = R_X(i-j)$$

Problème de la phase minimale

On rappelle l'approche utilisée pour notre problème de compression

- On se donne une séquence déterminée $x_0, \ldots, x_{N-1}$.
- On cherche les coefficients a_i qui minimisent le résidu ϵ. Ils sont donnés pas l'équation 7.1.
- On filtre les x_i par les coefficients a_i et on stocke les ϵ en même temps que les coefficients a_i (ou on les régénère comme un simple bruit blanc au moment du décodage.
- Au moment du décodage), on résout l'équation de récurrence, dont l'inconnue est la séquence y_i,

  $$a_0 y_n + a_1 y_{n-1} + \cdots + a_p y_{n-p} = \epsilon_n$$

  par l'étude faite au chapitre sur la TZ, cette équation n'a qu'une seule solution sommable. Or x l'est. Donc, si y est sommable $y = x$. On a reconstruit le signal d'origine. (si on a remplacé les vrais ϵ par un bruit blanc on espère que y ne sera trop différent d'un point de vue auditif du signal d'origine).

la dernière étape pose problème si le polynôme

$$P(z^{-1}) = \sum a_i z^{-i}$$

a des zéros (en z) de module plus grand que 1. En effet, dans ce cas là, l'implémentation récursive de l'équation de récurrence va renvoyer une séquence y_n explosive comme illustré au chapitre sur la TZ. (même si $y = x$ est la seule solution possible, en effet les seules erreurs d'arrondi suffiront à rendre impossible un calcul fiable)

Nous alors résoudre ce problème en remplaçant les coefficients a_i par d'autres qui garantiront la même énergie pour le résidu ϵ.

Notons $G(z)$ la TZ du résidu ϵ. Elle est donnée par

$$G(z) = X(z)P(z^{-1})$$

où $X(z)$ est la TZ de x. Elle est prescrite, car x est un signal fixé. Ainsi l'égalité de Parseval nous dit que la norme au carré de la suite ϵ est donnée par

$$\|\epsilon\|_2^2 = \int |X(e^{2i\pi\nu})|^2 |P(e^{-2i\pi\nu})|^2 d\nu.$$

Le module de ϵ ne change pas si P est remplacé par un polynôme qui a le même module en tout point du cercle unité.

Si z_0 est un zéro de $P(z^{-1})$ de module strictement supérieur à 1, alors la fonction définie sur le cercle unité par

$$z \mapsto P(z^{-1}) \frac{\overline{z_0} - z^{-1}}{1 - z_0 z^{-1}}$$

a le même module que $P(z^{-1})$ et est bien un polynôme en z^{-1}. On a remplacé le zéro z_0 par le zéro $\frac{1}{\overline{z_0}}$ qui est strictement à l'intérieur du cercle unité. En refaisant cette opération pour chaque zéro hors du disque unité, nous obtenons un polynôme P_1 qui a le même module que P sur le cercle unité, mais dont tous les zéros sont de module strictement inférieur 1. Ce sont les coefficients de ce polynôme qu'il faut prendre au lieu des a_i calculés. Ils nous fourniront un module des échantillons ϵ aussi petit que ce que fournissent les a_i avec l'avantage de retrouver le signal y $(=x)$ de manière stable.

7.4.2 Filtrage de Wiener

On suppose que l'on capte un processus SSL X. L'observation se fait au travers d'un canal de transmission qui filtre le processus X par un filtre sommable h et ajoute un bruit B (que l'on ne suppose pas forcément blanc). Le processus observé est nommé Y (il est SSL) et est donné par

$$Y = h * X + B$$

On suppose, pour simplifier, que B et X sont centrés et réels. On suppose aussi que le bruit est indépendant du signal. On suppose connues les DSP S_X et S_B. Par exemple, si X est un signal à variations lentes, sa DSP sera plus forte dans les dans les basses fréquences. Si B est supposé bruit blanc, alors on sait déjà que sa DSP sera une fonction constante sur $[-\frac{1}{2}, \frac{1}{2}[$.

On cherche un signal g tel que la puissance du processus $g * Y - X$ soit la plus faible possible.

Autrement dit, on cherche à définir un SLI dont la RI est g qui permettra, à partir du processus mesuré Y de retrouver au mieux le signal d'intérêt. Un tel SLI est appelé filtre de Wiener.

On appelle $Z = (g * Y) - X = ((g * h) - \delta) * X + g * B$. On voit tout de suite que si le bruit est nul il suffit de prendre pour g la RI du SLI inverse de celui défini par h, i.e. $g * h = \delta$, ce qui est conforme à l'intuition : En l'absence de bruit, il suffit d'inverser le SLI pour trouver le signal original.

D'après l'hypothèse d'indépendance entre X et B, on a

$$R_Z(k) = R_T(k) + (g * \tilde{g}) * R_B$$

où $T = ((g * h) - \delta) * X$ est un processus SSL.

et

$$S_Z(\nu) = |\hat{g}(\nu)|^2 S_B(\nu) + |\hat{g}(\nu)\hat{h}(\nu) - 1|^2 S_X(\nu)$$

la puissance du processus Z (qu'il faut minimiser) est donnée par (on rappelle qu'ici tous les processus sont centrés)

$$P_Z = \int S_Z(\nu) d\nu$$

Pour minimiser P_Z, il suffit de minimiser $S_Z(\nu)$ pour tout ν fixé. En effet, la seule variable inconnue est $\hat{g}(\nu)$ et elle n'intervient qu'une seule fois dans l'intégrale définissant P_Z. On choisissant $\hat{g}(\nu)$ pour chaque ν de manière à minimiser $S_Z(\nu)$ on n'interfère pas avec les autres valeurs de $S_Z(\nu')$ pour $\nu' \neq \nu$.

Fixons donc ν et pour alléger les notations. On note : $t = \hat{g}(\nu)$ (c'est l'inconnue), $\alpha = \hat{h}(\nu)$, $\rho_X = S_X(\nu)$ (c'est un réel positif), $\rho_B = S_B(\nu)$. On cherche t qui minimise

$$S_Z(\nu) = |t|^2 \rho_B + |t\alpha - 1|^2 \rho_X$$

La dernière expression se réécrit

$$S_Z(\nu) = |t|^2 \rho_B + \overline{(t\alpha - 1)}(t\alpha - 1)\rho_X = |t|^2(|\alpha|^2 \rho_X + \rho_B) - \overline{t\alpha}\rho_X - t\alpha\rho_X + \rho_X =$$

On sort ρ_X en facteur

$$\rho_X \left(|t|^2 \left(|\alpha|^2 + \frac{\rho_B}{\rho_X} \right) - \overline{t\alpha} - t\alpha \right) + \rho_X =$$

On note $\gamma = \left(|\alpha|^2 + \frac{\rho_B}{\rho_X} \right)$ qui est un réel et on continue

$$= \frac{\rho_X}{\gamma} \left(|t|^2 \gamma^2 - \overline{t\alpha}\gamma - t\alpha\gamma \right) + \rho_X = \frac{\rho_X}{\gamma} \left(\overline{(t\gamma - \overline{\alpha})}(t\gamma - \overline{\alpha}) - |\alpha|^2 \right) + \rho_X$$

$$= C\,|ty - \overline{\alpha}|^2 + D$$

Ou C et D sont des constantes qui dépendent de, $\rho_X,\ \gamma,\ \alpha$ mais pas de t. La valeur de t qui minimise la dernière équation est

$$t = \frac{\overline{\alpha}}{\gamma}$$

en réinterprétant en fonction des données du problème, on doit avoir

$$\forall \nu \in [-\frac{1}{2}, \frac{1}{2}[,\ \hat{g}(\nu) = \frac{\overline{\hat{h}(\nu)}}{|\hat{h}(\nu)|^2 + \frac{S_B(\nu)}{S_X(\nu)}}$$

Interprétation du résultat

Le résultat obtenu s'interprète comme suit :

- Lorsque le rapport $\frac{S_B(\nu)}{S_X(\nu)}$ est faible (ou nul si le bruit est absent) alors la formule de Wiener dit qu'il faut simplement inverser le filtre (diviser en fréquence par $\hat{h}(\nu)$).
- Lorsque le rapport $\frac{S_B(\nu)}{S_X(\nu)}$ est grand (ou infini si le signal ne porte pas de puissance à la fréquence ν) alors on met à zéro $\hat{g}(\nu)$. Cela est cohérent avec l'idée que pour cette fréquence, seul le bruit est intervenu et Y ne porte aucune information utile sur X à la fréquence ν.

Vous verrez en TP une autre application des outils introduits dans ce chapitre, il s'agira de résoudre l'équation suivante

$$Y = h * X + B$$

où Y représente un signal enregistré, X un signal émis, h la RI du canal de communication et B un bruit inconnu.

Par exemple, X peut être un signal connu de l'émetteur et du récepteur et h est inconnu. On a une seule équation et deux inconnues : h et B et pourtant, grâce aux outils introduits, on sera capable de la résoudre par une simple étude statistique. Le h recherché est celui qui minimisera l'énergie du processus $h * X - Y$.

Annexe A

La transformation de Fourier (pour $\mathbb{R}$ et $[-\frac{1}{2}, \frac{1}{2}[$)

Dans ce chapitre[1] nous introduisons la Transformée de Fourier sur $\mathbb{R}$ et $[-\frac{1}{2}, \frac{1}{2}[$. Ces deux transformées de Fourier sont utiles pour le chapitre suivant qui porte sur l'échantillonnage des signaux définis sur $\mathbb{R}$.

Ce chapitre sera très formel. Le lecteur pourra trouver les démonstrations dans divers documents présents sur le site pédagogique. Ce qu'il faut en retenir est que les règles formelles autour de toute transformation de Fourier (comme, par exemple, le fait que la convolution se transforme en multiplication), sont applicables dès que les règles de calcul (que nous voyons en début de chapitre) le permettent.

A.1 Transformation de Fourier sur $\mathbb{R}$, ou encore Transformation de Fourier à Temps Continu(TFTC)

A.1.1 Espaces fonctionnels et règles de calcul

On définit, comme on l'a fait sur $\mathbb{Z}$, les trois espaces fonctionnels suivants

Définition A.1. Espaces de fonctions
On définit les trois espaces de fonctions définies sur $\mathbb{R}$ principaux suivants :

1. *On note $L^1(\mathbb{R})$ l'espace des fonctions* **sommables** *c'est-à-dire des fonctions f qui vérifient :*

$$\int_{\mathbb{R}} |f(x)|dx < +\infty.$$

on note

$$\|f\|_1 = \int_{\mathbb{R}} |f(x)|dx$$

C'est une norme sur l'espace des fonctions sommables.

1. Ce qui est dit dans ce chapitre a déjà été vu en cours de mathématiques (MDI103). Il est inclus pour référence

2. *On note* $L^2(\mathbb{R})$ *l'espace des fonctions d'***énergie finie** *c'est-à-dire des fonctions* f *qui vérifient :*

$$\int_{\mathbb{R}} |f(x)|^2 dx < +\infty.$$

et on note

$$\|f\|_2 = \left(\int_{\mathbb{R}} |f(x)|^2 dx \right)^{\frac{1}{2}}$$

Il s'agit d'une norme sur l'espace des fonctions d'énergie finie.

3. *On note* $L^\infty(\mathbb{R})$ *l'espace des fonctions* **bornées** *c'est-à-dire des fonctions* f *qui vérifient :*

$$\exists C \in \mathbb{R}_+, \quad \textit{tel que pour presque tout (p.t) } x \in \mathbb{R}, \ |f(x)| \leqslant C.$$

et on note

$$\|f\|_\infty = \inf\{C : \ \textit{vérifie l'équation ci-dessus}\}$$

c'est une norme sur l'espace des fonctions bornées.

Proposition A.2. Quelques cas d'inclusion
Pour les suites nous avions la propriété suivante

$$l^1 \subset l^2 \subset l^\infty$$

Sur $\mathbb{R}$ *nous n'avons pas ces inclusions. Nous avons par contre les propriétés suivantes*

1. *Si* $f \in L^1(\mathbb{R})$ *et* $f \in L^\infty(\mathbb{R})$ *alors* $f \in L^2(\mathbb{R})$.
2. *Si* $f \in L^2(\mathbb{R})$ *et* $\exists A > 0, \ t.q \ , \ (|x| \geqslant A) \Longrightarrow f(x) = 0$ *(on dit que* f *est à support borné). Alors* $f \in L^1(\mathbb{R})$

On peut retenir ces règles sous la forme :
Pour une fonction à support borné on a, (bornée$\Longrightarrow$*d'énergie finie*$\Longrightarrow$*sommable).*
Pour une fonction bornée on a, (sommable$\Longrightarrow$*d'énergie finie). Par exemple dans les espaces de suites* l^1, l^2, l^∞ *les suites sont bornées, d'où la propriété par laquelle nous avons commencé cette proposition.*

Démonstration. □

1. Si $f \in L^1(\mathbb{R})$ et $f \in L^\infty(\mathbb{R})$ alors

$$\int |f|^2 \leqslant \int \|f\|_\infty |f| = \|f\|_\infty \|f\|_1$$

2. Si f est d'énergie finie et à support fini alors

$$\int |f| = \int_{|f(x)|>1} |f(x)|dx + \int_{0<|f(x)|\leqslant 1} |f(x)|dx + \int_{f(x)=0} |f(x)|dx$$

$$\leqslant \int_{|f(x)|>1} |f(x)|^2 dx + \int_{0<|f(x)|\leqslant 1} 1dx + 0 \leqslant \|f\|_2^2 + 2A < +\infty$$

Proposition A.3. Règles de calcul de convolution et de produit
Nous avons les règles suivantes pour la multiplication et le produit de convolution

1. *Si* $(f,g) \in L^k(\mathbb{R}) \times L^\infty(\mathbb{R})$ *(*k *vaut 1, 2 ou* ∞*) alors*

$$fg \in L^k(\mathbb{R}) \text{ et } \|fg\|_k \leqslant \|f\|_k \|g\|_\infty$$

2. *Si* $(f,g) \in L^k(\mathbb{R}) \times L^1(\mathbb{R})$ *(*k *vaut 1, 2 ou* ∞*) alors*

$$f * g \in L^k(\mathbb{R}) \text{ et } \|f * g\|_k \leqslant \|f\|_k \|g\|_1$$

3. *Si* $f, g \in L^2(\mathbb{R})$ *alors*

   $f * g \in L^\infty(\mathbb{R})$ *et* $f * g$ *est continue et tend vers 0 à l'infini et* $\|f * g\|_\infty \leqslant \|f\|_2 \|g\|_2$

   et

$$fg \in L^1(\mathbb{R}) \text{ et } \|fg\|_1 \leqslant \|f\|_2 \|f\|_2$$

   Cette dernière inégalité est aussi appelée inégalité de Cauchy-Schwartz.

A.1.2 Définition et propriétés habituelles

Définition A.4. Transformée de Fourier pour les fonctions de $\mathbb{R}$, ou Transformée de Fourier à Temps Continu (TFtC)
Pour toute fonction $f \in L^1(\mathbb{R})$ *on définit sa transformée de Fourier à temps continu que l'on note* $\hat{f}$ *ou* $\mathcal{F}(f)$*, par*

$$\forall \nu \in \mathbb{R},\ [\mathcal{F}(f)](\nu) = \hat{f}(\nu) = \int f(x) e^{-2i\pi\nu x} dx$$

Et dans ce cas, $\hat{f}$ *est continue et tend vers 0 à l'infini et*

$$\|\hat{f}\|_\infty \leqslant \|f\|_1$$

Exemple A.5.
Par exemple si $f = \mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}$ est l'indicatrice de l'intervalle $[-\frac{1}{2},\frac{1}{2}]$ alors sa TFtC est la fonction sinus cardinal

$$\hat{f}(\nu) = \frac{\sin(\pi\nu)}{\pi\nu} = \mathrm{sinC}(\pi\nu)$$

($\mathrm{sinC}(x)$ et le rapport entre $\sin(x)$ et x lorsque x est non nul et $\mathrm{sinC}(0) = 1$).

Cette fonction est bien continue et tend vers zéro à l'infini.

On a les propriétés suivantes

Proposition A.6. Propriétés de la TFtC
Dans la suite f *et* g *sont des fonctions de* $L^1(\mathbb{R})$ *et* ν_0 *un élément de* $\mathbb{R}$ *et* $\varphi(x) = e^{2i\pi\nu_0 x}$ *est une onde de Fourier sur* $\mathbb{R}$ *de fréquence* ν_0*.*

1. *La convolution est transformée en produit :*

$$\mathcal{F}(f * g) = \hat{f}\hat{g}$$

2. *Le produit (de deux transformées) est transformé en convolution si les règles de calcul le permettent, par exemple si* $\hat{f} \in L^1(\mathbb{R})$ *:*

$$\overline{\mathcal{F}}(\hat{f}\hat{g}) = f * g.$$

(L'opérateur $\overline{\mathcal{F}}$ *consiste à effectuer une TFtC et à symétriser le résultat).*

3. *Multiplier par une onde de Fourier de fréquence* ν_0 *revient à décaler la transformée de* ν_0 *:*

$$\forall \nu \in \mathbb{R},\ [\mathcal{F}(\varphi.f)](\nu) = \hat{f}(\nu - \nu_0)$$

4. *Un décalage de la fonction de y revient à multiplier la TFtC par une onde de Fourier de fréquence* $-y$*. On note* f_y *la y-translatée de* f *(*$f_y(x) = f(x-y)$*) :*

$$\widehat{f_y}(\nu) = \hat{f}(\nu)e^{-2i\pi y\nu}$$

5. *Si* f *est réelle, alors* $\hat{f}$ *possède la* **symétrie hermitienne** *:*

$$(\forall x \in \mathbb{R},\ f(x) \in \mathbb{R}) \Longrightarrow \Big(\ \forall \nu \in \mathbb{R},\ \hat{f}(-\nu) = \overline{\hat{f}(\nu)} \qquad \Big)$$

6. *Si* f *est symétrique alors* $\hat{f}$ *aussi :*

$$(\forall x \in \mathbb{R},\ f(-x) = f(x)) \Longrightarrow \Big(\ \forall \nu \in \mathbb{R},\ \hat{f}(-\nu) = \hat{f}(\nu) \qquad \Big)$$

7. *Si* f *est à la fois symétrique et réelle alors* $\hat{f}$ *est aussi symétrique et réelle (cela se déduit des deux propriétés précédentes).*

Proposition A.7. Re-normalisation du temps
Si $f \in L^1(\mathbb{R})$ *et que* $\lambda > 0$ *est un réel alors*

$$\mathcal{F}\left[x \mapsto f(\lambda x)\right](\nu) = \frac{1}{\lambda}\hat{f}\left(\frac{\nu}{\lambda}\right)$$

Lorsque λ *est plus grand que 1, la fonction* $x \mapsto f(\lambda x)$ *est un rétrécissement de* f *autour de l'axe des ordonnée, et dans ce cas la transformée de Fourier est zommée autour de l'axe des ordonnées (et multipliée par* $1/\lambda$*).*

Démonstration.

Faire le changement de variable $y = \lambda x$ dans la formule de calcul de la transformée de Fourier. □

Exemple A.8.
La transformée de Fourier la fonction $\mathbb{1}_{[-A/2,A/2]}$ (indicatrice de l'intervalle $[-A/2, A/2]$) est la fonction $\nu \mapsto A\,\mathrm{sinC}(A\pi\nu)$. En effet, on utilise la proposition ci-dessus en remarquant que

$$\mathbb{1}_{[-\frac{A}{2},\frac{A}{2}]}(t) = \mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}\left(\frac{t}{A}\right) \text{ i.e. } \lambda = \frac{1}{A}$$

A.1.3 Théorème d'inversion

On définit la transformation de Fourier inverse. En fait, elle est très proche de la transformée de Fourier, les deux transformées ne diffèrent entre elle que d'un signe dans l'équation de l'onde de Fourier, elle sera notée $\overline{\mathcal{F}}$.

Définition A.9. la transformée de Fourier inverse
Si f est une fonction de $L^1(\mathbb{R})$ on note $\overline{\mathcal{F}}(f)$ la transformée de Fourier inverse de f. Elle est définie par l'équation suivante :

$$\forall \nu \in \mathbb{R},\ [\overline{\mathcal{F}}(f)](\nu) = \int_{t\in\mathbb{R}} f(t)e^{+2i\pi\nu t}dt$$

On remarque que $[\overline{\mathcal{F}}(f)](-\nu) = [\mathcal{F}(f)](\nu)$.

On a le théorème suivant (non démonté ici)

Théorème A.10. Théorème d'inversion
Si $f \in L^1(\mathbb{R})$ et $\hat{f} \in L^1(\mathbb{R})$ alors on a

$$\overline{\mathcal{F}}\left(\mathcal{F}(f)\right) = f$$

ou en version ponctuelle[2]

$$\forall x \in \mathbb{R},\ f(x) = \int \hat{f}(t)e^{2i\pi xt}dt$$

On en déduit, en particulier, que si $\hat{f}$ est sommable alors f est continue et de limite nulle à l'infini car on a pu l'écrire comme une transformée de Fourier d'une fonction sommable.

On a aussi, par un raisonnement symétrique

$$f = \mathcal{F}\left(\overline{\mathcal{F}}(f)\right)$$

Exemple A.11.
Ce théorème permet de dire que sinC n'est pas une fonction sommable, car elle est la TF d'une fonction sommable non continue. (évidemment, on peut arriver à cette conclusion de manière plus simple).

Corollaire A.12. Injectivité de la TFtC
Le corollaire du théorème d'inversion est que $\mathcal{F}$ est injective.

En effet, si $\mathcal{F}(f) = 0$ alors $\mathcal{F}(f)$ est bien $L^1(\mathbb{R})$ (parce que nulle) et le théorème d'inversion s'applique et donne $f = 0$, ce qui est la caractérisation de l'injectivité des applications linéaires.

Exemple A.13. **Untilisation pour démontrer une propriété générale**
Nous avons vu que si f et g sont sommables alors $\mathcal{F}(f * g) = \hat{f}\hat{g}$. Si on suppose en plus que $\hat{f}$ est sommable alors $\hat{f}\hat{g}$ est aussi sommable par les règles de calcul et le théorème d'inversion permet de dire que

$$f * g = \overline{\mathcal{F}}(\hat{f}\hat{g})$$

(c'était la propriété 2 de la proposition A.6)

2. En toute rigueur, il faudrait dire que f possède un représentant dans sa classe d'équivalence de l'"égalité presque partout" qui vérifie l'égalité ponctuelle.

Exemple A.14. **Exemple d'utilisation pour le calcul d'une transformée de Fourier**
Nous voulons calculer la transformée de Fourier de la fonction g définie par

$$g(x) = \left(\frac{\sin(\pi x)}{\pi x}\right)^2 = (\mathrm{sinC}(\pi x))^2$$

Poser la formule intégrale du calcul de la transformée de Fourier ne débouche pas sur un calcul immédiat.

Comme vu plus haut la TF de l'indicatrice de $[-\frac{1}{2}, \frac{1}{2}]$ est

$$\nu \mapsto \mathrm{sinC}(\pi\nu)$$

Nous remarquons que la transformée de Fourier de la fonction (fonction triangle)

$$f(x) = \begin{cases} 1 - |x| \text{ si } |x| < 1 \\ 0 \text{ sinon} \end{cases}$$

est précisément la fonction g. En effet, f est la convolée de l'indicatrice de $[-\frac{1}{2}, \frac{1}{2}]$ avec elle-même (faites le calcul). Et comme la convolution est transformée en multiplication par $\mathcal{F}$, on en déduit que la TF de f est le carré de celle du créneau, c'est à dire g.

Or, g est une fonction de $L^1(\mathbb{R})$ (elle décroît comme $1/x^2$) et donc le théorème d'inversion s'applique et donne

$$\forall x \in \mathbb{R},\ f(x) = [\overline{\mathcal{F}}(g)](x) = \int g(t)e^{2i\pi xt}dt$$

Le changement de variable $u = -t$ donne

$$\forall x \in \mathbb{R},\ f(x) = \int g(u)e^{-2i\pi xu}du \text{ (g est une fonction paire)}$$

Donc la transformée de Fourier de g est la fonction f.

A.1.4 Extension à $L^2(\mathbb{R})$

Si une fonction est d'énergie finie ($\in L^2(\mathbb{R})$) la définition par intégrale de la TF ne permet *a priori* pas de calculer une transformée de Fourier pour une telle fonction. Cependant, nous avons le théorème suivant dont le point central est l'égalité de Parseval.

Théorème A.15. *Il existe une unique application, que nous notons provisoirement $\mathcal{F}_2$, linéaire de $L^2(\mathbb{R})$ vers $L^2(\mathbb{R})$ qui a les propriétés suivantes*

1. *Elle coïncide avec la transformée de Fourier sur $L^1(\mathbb{R})$*

$$\forall f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}),\ \mathcal{F}_2(f) = \mathcal{F}(f)$$

   (ceci implique en particulier que si $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ alors $\mathcal{F}(f) \in L^2(\mathbb{R})$)

2. *C'est une isométrie (égalité de Parseval)*

$$\forall f \in L^2(\mathbb{R}),\ \|\mathcal{F}_2(f)\|_2 = \|f\|_2$$

3. *C'est une bijection entre $L^2(\mathbb{R})$ et lui-même. En particulier $\forall g \in L^2(\mathbb{R}),\ \exists f \in L^2(\mathbb{R}),\ t.q.\ \mathcal{F}_2(f) = g$ (c'est la surjectivité).*

Démonstration.
Les étapes de la démonstration (qui n'est pas faite ici) sont

1. Remarquer que les fonctions $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ sont denses dans $L^1(\mathbb{R})$.
2. Donc il ne peut y avoir qu'une seule extension continue de $\mathcal{F}$ de $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ à $L^2(\mathbb{R})$.
3. Démontrer l'égalité de Parseval pour les fonctions de $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$.
4. Définir $\mathcal{F}_2(f)$ comme la limite (unique) dans $L^2(\mathbb{R})$ de $\mathcal{F}(f_n)$ où f_n est une suite de fonctions qui tendent vers f pour la norme $\|.\|_2$.
5. Montrer qu'une telle extension est linéaire et isométrique (facile)
6. Montrer la surjectivité. (l'injectivite de $\mathcal{F}_2$ vient du fait que c'est une isométrie).

□

Théorème A.16. *Théorème d'inversion pour $L^2(\mathbb{R})$*
La fonction $\overline{\mathcal{F}}$ peut être étendue de la même manière à $L^2(\mathbb{R})$ et on note provisoirement son extension $\overline{\mathcal{F}_2}$. On a

$$\forall f \in L^2(\mathbb{R}),\ \overline{\mathcal{F}_2}(\mathcal{F}_2(f)) = \mathcal{F}_2(\overline{\mathcal{F}_2}(f)) = f$$

i.e le théorème d'inversion pour $L^2(\mathbb{R})$ ne fait aucune hypothèse sur $\hat{f}$ pour s'appliquer

Remarque A.17. **Comment calculer concrètement la TF d'une fonction $L^2(\mathbb{R})$?**
Les deux derniers théorèmes permettent de donner un sens à la TF d'une fonction d'énergie finie même si celle-ci n'est pas sommable. Cependant, la définition de $\mathcal{F}_2$ n'est pas très explicite. Comment allons-nous calculer la transformée de Fourier pour une fonction $f \in L^2(\mathbb{R})$? Plusieurs cas peuvent survenir, par exemple :

1. $f \in L^1(\mathbb{R})$: Si f est intégrable, alors par définition de $\mathcal{F}_2$ on a

   $$\mathcal{F}_2(f) = \mathcal{F}(f)$$

   On peut donc obtenir $\mathcal{F}_2(f)$ par un calcul d'intégrale.
2. Il se trouve que l'on connaît une fonction $L^1(\mathbb{R})$ dont f est la transformée de Fourier : Par exemple,

   $$f(x) = \mathrm{sinC}(\pi x) = \frac{\sin(\pi x)}{\pi x}$$

   Notons $g = \mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}$ l'indicatrice de $[-\frac{1}{2},\frac{1}{2}]$. On sait que

   $$\mathcal{F}(g) = f$$

   Et comme g est dans $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ on a aussi

   $$\mathcal{F}_2(g) = f$$

   Et par symétrie (f est symétrique)

   $$\overline{\mathcal{F}_2}(g) = f$$

On a donc par théorème d'inversion

$$g = \mathcal{F}_2(\overline{\mathcal{F}_2}(g)) = \mathcal{F}_2(f)$$

Ainsi la transformée de Fourier du sinus cardinal (fonction f ici) est le créneau indicateur de $[-\frac{1}{2}, \frac{1}{2}]$.

3. Dans le cas général : On sait que si $f \in L^2(\mathbb{R})$ alors les fonctions f_N définies par

$$f_N(x) = \begin{cases} f(x) \text{ si } |x| < N \\ 0 \text{ sinon} \end{cases}$$

sont dans $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ (car elles sont d'énergie finie comme restriction de f à un intervalle et sommables car d'énergie finie et à support fini). On sait aussi que $\|f_N - f\|_2$ tend vers 0 lorsque N tend vers l'infini (théorème de convergence dominée). On en déduit que $\mathcal{F}_2(f_N)$ approche $\mathcal{F}_2(f)$. Or $\mathcal{F}_2(f_N)$ peut être obtenu par un calcul intégral. Toute autre type de suite tendant vers f peut être utilisée.

À partir de maintenant nous ne notons plus $\mathcal{F}_2$ mais simplement $\mathcal{F}$, il ne peut y avoir de problème de confusion car les deux opérateurs coïcindent si la fonction est à la fois $L^1(\mathbb{R})$ et $L^2(\mathbb{R})$ et si elle est seulement $L^1(\mathbb{R})$ ou $L^2(\mathbb{R})$, une seule des deux notations $\mathcal{F}_2(f)$ ou $\mathcal{F}(f)$ a un sens.

Quelles sont les propriétés de la TF sur $L^2(\mathbb{R})$? Nous nous bornons à énoncer un principe qui vous permettra de retrouver ces propriétés

> Toutes les propriétés formelles données dans la proposition A.6 sont vraies en remplaçant f et g par des fonctions $L^2(\mathbb{R})$ ou $L^1(\mathbb{R})$ **à partir du mement où les règles de calcul donnent un sens à l'un des termes de l'égalité**

Exemple A.18. Voici des exemples d'utilisation du principe ci-dessus :

1. Si f et g sont dans $L^2(\mathbb{R})$, a-t-on

$$\mathcal{F}(f * g) = \hat{f}.\hat{g}?$$

Les règles de calcul nous disent sur $f * g$ est seulement $L^\infty(\mathbb{R})$ le terme de gauche est donc sans objet (nous n'avons défini la TF que pour $L^1(\mathbb{R})$ et $L^2(\mathbb{R})$) et cette égalité est, à priori, **fausse**.

2. Autre exemple : toujours avec f et g dans $L^2(\mathbb{R})$ a-t-on

$$f * g = \overline{\mathcal{F}}(\hat{f}.\hat{g})?$$

Le terme de droite est la TF (inverse) du produit de $\hat{g}$ et $\hat{f}$ qui sont deux fonctions $L^2(\mathbb{R})$. Ce produit est donc, par les règles de calcul, $L^1(\mathbb{R})$. Le terme de droite a donc un sens. Le terme de gauche a aussi un sens, $f * g$ est une fonction continue, bornée et tendant vers 0 à l'infini (règle de produit de convolution $L^2(\mathbb{R})$ contre $L^2(\mathbb{R})$) et **oui, cette égalité est vraie** sous l'hypothèse $f, g \in L^2(\mathbb{R})$

3. Encore une fois : Si $f \in L^1(\mathbb{R})$ et $g \in L^2(\mathbb{R})$, a-t-on

$$\mathcal{F}(f * g) = \hat{f}.\hat{g}$$

oui, car la convolée de f et g est d'énergie finie. Donc le terme de gauche a un sens. Le terme de droite est aussi d'énergie finie (produit de la fonction bornée (et continue) $\hat{f}$ par la fonction d'énergie finie $\hat{g}$)

Nous réécrivons dans la proposition suivante les propriétés habituelles de la TF en les reformulant dans la cas $L^2(\mathbb{R})$. Vous pourrez vérifier qu'il s'agit d'une réécriture de la proposition A.6 en utilisant notre principe encadré ci-sessus.

Proposition A.19. Propriétés de la TFtC
Dans la suite f et g sont des fonctions de $L^2(\mathbb{R})$ et ν_0 un élément de $\mathbb{R}$ et $\varphi(x) = e^{2i\pi\nu_0 x}$ est une onde de Fourier sur $\mathbb{R}$ de fréquence ν_0.

1. *La convolution de deux fonctions s'exprime comme une TF (inverse) du produit des transformées :*
$$f * g = \overline{\mathcal{F}}(\hat{f}\hat{g})$$
2. *Le produit est transformé en convolution :*
$$\mathcal{F}(fg) = \hat{f} * \hat{g}$$
(Ici fg est une fonction sommable et $\mathcal{F}(fg)$ est la TF au sens $L^1(\mathbb{R})$)
3. *Multiplier par une onde de Fourier de fréquence ν_0 revient à décaler la transformée de ν_0 :*
$$\forall \nu \in \mathbb{R},\ [\mathcal{F}(\varphi.f)](\nu) = \hat{f}(\nu - \nu_0)$$
4. *Un décalage de la fonction de y revient à multiplier la TFtC par une onde de Fourier de fréquence $-y$. On note f_y la y-translatée de f ($f_y(x) = f(x-y)$) :*
$$\widehat{f_y}(\nu) = \hat{f}(\nu)e^{-2i\pi y\nu}$$
5. *Si f est réelle, alors $\hat{f}$ possède la* **symétrie hermitienne** *:*
$$(\forall x \in \mathbb{R},\ f(x) \in \mathbb{R}) \Longrightarrow \Big(\ \forall \nu \in \mathbb{R},\ \hat{f}(-\nu) = \overline{\hat{f}(\nu)} \quad \Big)$$
6. *Si f est symétrique alors $\hat{f}$ aussi :*
$$(\forall x \in \mathbb{R},\ f(-x) = f(x)) \Longrightarrow \Big(\ \forall \nu \in \mathbb{R},\ \hat{f}(-\nu) = \hat{f}(\nu) \quad \Big)$$
7. *Si f est à la fois symétrique et réelle alors $\hat{f}$ est aussi symétrique et réelle (cela se déduit des deux propriétés précédentes).*

A.1.5 Échange de régularité de décroissance à l'infini

Nous avons le résultat suivant (qui est aussi général à toutes les transformations de Fourier). Il se résume en disant que plus une fonction est régulière, plus sa TF décroît rapidement à l'infini. Et plus une fonction décroît rapidement à l'infini, plus sa transformée de Fourier est régulière.

Théorème A.20. *Échange de régularité et de décroissance à l'infini*
Si f est une fonction $L^1(\mathbb{R})$ alors

1. *Si*
$$\forall 0 \leqslant k \leqslant K,\ (x \mapsto x^k f(x)) \in L^1(\mathbb{R})$$
(i.e. la fonction f décroît assez vite pour que la multiplication par un polynôme ne suffise pas à la rendre non sommable). Alors $\hat{f}$ est K-fois continûment dérivable et
$$(\mathcal{F}(f))^{(k)} = \mathcal{F}(x \mapsto (-2i\pi x)^k f(x))$$
En particulier si $K = \infty$, alors $\hat{f}$ est $\mathcal{C}^\infty$.
2. *Si f est K-fois continûment dérivable est que toutes ses dérivées sont $L^1(\mathbb{R})$ alors*
$$\forall 0 \leqslant k \leqslant K,\ \lim_{|\nu \to +\infty} \nu^k \hat{f}(\nu) = 0$$
et la TF de la dérivée k-ième de f est (pour $k \leqslant K$)
$$\mathcal{F}(f^{(k)})(\nu) = (2i\pi\nu)^k \hat{f}(\nu)$$
en particulier, si f est $\mathcal{C}^\infty$ et que toutes ses dérivées sont sommables, sa TF décroît plus vite à l'infini que n'importe quel inverse de polynôme.

Exemple A.21.
La fonction indicatrice de l'intervalle $[-\frac{1}{2}, \frac{1}{2}]$ décroît extrêmement rapidement à l'infini (elle est nulle au-delà de 1/2) et sa TF qui est la fonction $\mathrm{sinC}(\pi x)$ est effectivement extrêmement régulière. Non seulement sinC est $\mathcal{C}^\infty$ mais de plus sa valeur en n'importe quel point est donnée par
$$\forall x \in \mathbb{R},\ \mathrm{sinC}(x) = \sum_{n \geqslant 0} (-1)^n \frac{x^{2n+1}}{(2n+1)!}$$
qui signifie que cette fonction est entière (une classe de fonctions que l'on peut connaître à partir de leur connaissance sur n'importe quel intervalle de taille non nulle).

Par contre la fonction créneau n'est pas du tout régulière et on constate bien que la fonction sinC décroît lentement à l'infini (comme $1/x$).

A.2 Transformation de Fourier pour $[-\frac{1}{2}, \frac{1}{2}[$, ou coefficients de Fourier

Dans cette partie nous nous contentons de la définition des coefficients de Fourier d'une fonction définie sur $[-\frac{1}{2}, \frac{1}{2}[$ et de quelques propriétés utiles pour le chapitre suivant. Le lecteur pourra utiliser le théorème d'inversion pour déduire toutes les propriétés de cette partie depuis celle que nous avons déjà exposées pour la TFtD.

On note $L^1([-\frac{1}{2}, \frac{1}{2}[)$, $L^2([-\frac{1}{2}, \frac{1}{2}[)$ et $L^\infty([-\frac{1}{2}, \frac{1}{2}[)$ les espaces de fonctions définies sur $[-\frac{1}{2}, \frac{1}{2}[$ qui sont sommables, d'énergie finie et bornées (respectivement).

On a les inclusions suivantes
$$L^\infty([-\frac{1}{2}, \frac{1}{2}[) \subset L^2([-\frac{1}{2}, \frac{1}{2}[) \subset L^1([-\frac{1}{2}, \frac{1}{2}[)$$

Définition A.22. Transformée de Fourier des fonctions périodiques ou "coefficients de Fourier"
Si $f \in L^1([-\frac{1}{2}, \frac{1}{2}[)$ est une fonction sommable. On appelle coefficients de Fourier de f les nombres complexes c_k définis par

$$\forall k \in \mathbb{Z},\ c_k = \int_{-\frac{1}{2}}^{\frac{1}{2}} f(t)e^{-2i\pi kt}dt$$

Les c_k tendent vers 0 lorsque $|k|$ tend vers l'infini.

En raison de l'inclusion $L^2([-\frac{1}{2}, \frac{1}{2}[) \subset L^1([-\frac{1}{2}, \frac{1}{2}[)$, l'extension aux fonctions d'énergie finie sur $[-\frac{1}{2}, \frac{1}{2}[$ de cette transformée de Fourier est triviale (la définition pour $L^1([-\frac{1}{2}, \frac{1}{2}[)$ suffit).

On a les propriétés suivantes

Proposition A.23. Propriétés de la transformée de fourier des fonctions périodiques
Dans la suite f et g sont deux fonctions sur $[-\frac{1}{2}, \frac{1}{2}[$ (on précisera à chaque fois si on les suppose sommable ou d'énergie finie). Les c_k sont les coefficients de Fourier de f et les d_k ceux de g.

1. *La suite des coefficients de Fourier de l'onde $t \mapsto e^{2i\pi kt}$ est l'impulsion en en k, δ_n^k.*
2. *Si f et g sont sommables, et que les l_k sont les coefficients de Fourier de $f * g$ alors :*
$$l = cd$$
3. *Si les l_k sont les coefficients de Fourier de fg (toutes deux $L^2([-\frac{1}{2}, \frac{1}{2}[)$, ou une bornée et l'autre $L^1([-\frac{1}{2}, \frac{1}{2}[)$) alors*
$$l = c * d$$
4. *Multiplier par une onde de Fourier de fréquence k_0 revient à décaler les coefficients de k_0. Soit, si $g(t) = f(t)e^{2i\pi k_0 t}$ alors*
$$d_k = c_{k-k_0}$$
5. *Un décalage de la fonction de y revient à multiplier les coefficients de Fourier par une onde de Fourier de fréquence $-y$. Si $g = f_y$ où f_y la y-translatée de f ($f_y(x) = f(x-y)$, translation circulaire, voir chapitre 1) :*
$$d_k = c_k e^{-2i\pi yk}$$
6. *Si f est réelle, alors les c_k sont à symétrie hermitienne :*
$$c_{-k} = \overline{c_k}$$
7. *Si f est paire alors les c_k aussi, $c_{-k} = c_k$.*
8. *Si f est à la fois paire et réelle alors les c_k sont aussi symétriques et réels (cela se déduit des deux propriétés précédentes).*

On a encore l'égalité de Parseval.

Proposition A.24. Égalité de Parseval
Si $f \in L^2([-\frac{1}{2}, \frac{1}{2}[)$ et que les c_k sont ses coefficients de Fourier alors, la suite c_k est dans l'espace l^2 et

$$\int |f|^2 = \sum |c_k|^2$$

On a le théorème d'inversion (que nous déclinons pour le cas $L^1([-\frac{1}{2}, \frac{1}{2}[)$ et $L^2([-\frac{1}{2}, \frac{1}{2}[)$). Ce théorème est le pendant de celui vu au chapitre portant sur la TFtD.

Théorème A.25. Théorème d'inversion
Si f est une fonction de $L^1([-\frac{1}{2}, \frac{1}{2}[)$ et que les c_k sont ses coefficients de Fourier alors

1. *Si la suite des c_k est sommable (i.e. $\sum |c_k| < +\infty$) alors*

$$\forall t \in [-\frac{1}{2}, \frac{1}{2}[,\ f(t) = \sum_{k\in\mathbb{Z}} c_k e^{2i\pi tk}$$

   Ce qui signifie que $t \mapsto f(-t)$ est la TFtD de la suite c_k.

2. *Si f est dans $L^2([-\frac{1}{2}, \frac{1}{2}[)$ (dont on rappelle que c'est un sous-ensemble de $L^1([-\frac{1}{2}, \frac{1}{2}[)$) alors la suite de fonctions (indexée par K)*

$$t \mapsto \sum_{k=-K}^{k=K} c_k e^{2i\pi kt}$$

   tend en norme 2 vers f, c'est -à-dire que

$$\lim_{K\to+\infty} \int \left| f(t) - \sum_{k=-K}^{k=K} c_k e^{2i\pi kt} \right|^2 dt = 0$$

Enfin, si la fonction f est régulière, ses coefficients de Fourier décroissent vite à l'infini.

Proposition A.26. La régularité devient décroissance des coefficients
Si $f \in L^1([-\frac{1}{2}, \frac{1}{2}[)$ et que les c_k sont ses coefficients de Fourier. Alors si f est N-fois continûment dérivable (souvenez-vous que la continuité sur $[-\frac{1}{2}, \frac{1}{2}[$ implique d'admettre une limite en $1/2$ qui est égale à la valeur en $-1/2$) alors

$$\lim_{|k|\to+\infty} k^N c_k = 0$$

et les coefficients de Fourier de $f^{(l)}$ sont les

$$(2i\pi k)^l c_k$$

Remarque A.27. **Retour sur la transformée en Z**
Dans le chapitre sur la TZ nous avons manipulé des suites dont la TZ était une fraction rationnelle. Les fractions rationnelles sont des fonctions extrêmement régulières et nous avons vu que les suites dont elles sont la TZ (c'est-à-dire la TFtD définie sur le cercle unité par le changement de variable $\nu \mapsto e^{2i\pi\nu}$) sont à décroissance exponentielle. Cela est cohérent avec la dernière proposition ainsi qu'avec son pendant dans le chapitre TFtD (on avait vu que la TFtD d'une suite qui décroît rapidement à l'infini est une fonction régulière).

A.3 Coefficients de Fourier des fonctions définies sur $[-\frac{A}{2}, \frac{A}{2}[$ (renormalisation du temps)

Les fonctions définies sur $[-\frac{1}{2}, \frac{1}{2}[$ représentent les fonctions périodiques de période 1. Leur transformée de Fourier (que nous appelons coefficients de Fourier) est définie sur $\mathbb{Z}$. On peut aussi s'intéresser aux fonctions périodiques de période A. Les ondes de Fourier sur $[-\frac{1}{2}, \frac{1}{2}[$ avaient comme expression

$$\forall t \in [-\frac{1}{2}, \frac{1}{2}[,\ t \mapsto e^{2i\pi nt}$$

(c'est l'onde de fréquence $n \in \mathbb{Z}$).

Les ondes de Fourier sur $[-\frac{A}{2}, \frac{A}{2}[$ ont pour équation

$$\forall t \in [-\frac{A}{2}, \frac{A}{2}[,\ t \mapsto e^{2i\pi n \frac{t}{A}}$$

C'est l'onde de fréquence $\frac{n}{A}$.

On définit les coefficients de Fourier d'une fonction f définie sur $[-\frac{A}{2}, \frac{A}{2}[$ par

$$\forall k \in \mathbb{Z},\ c_k = \frac{1}{A}\int_{-\frac{A}{2}}^{\frac{A}{2}} f(t)e^{-2i\pi k\frac{t}{A}}dt$$

Le coefficient de normalisation $\frac{1}{A}$ est là pour faire que la suite des coefficients de Fourier d'une onde soit une impulsion. Les propriétés des coefficients sont les mêmes que dans le cas de $[-\frac{1}{2}, \frac{1}{2}[$. Par contre l'égalité de Parseval est modifiée de la manière suivante : Si $f \in \mathrm{L}^2([-\frac{A}{2}, \frac{A}{2}[)$ et que les c_k sont ses coefficients de Fourier alors

$$\int_{-\frac{A}{2}}^{\frac{A}{2}} |f(t)|^2 dt = A\sum_{k\in\mathbb{Z}} |c_k|^2$$

soit

$$\|f\|_2 = \sqrt{A}\|c\|_2$$

Remarque A.28. La question de la "bonne" normalisation de la transformation de Fourier ne peut pas être tranchée. Certains préfèrent utiliser une normalisation qui fait que l'égalité de Parseval soit préservée, par contre, dans ce cas, la TF d'une onde de Fourier ne sera pas une impulsion. Évidemment, le changement de normalisation ne change absolument rien aux propriétés profondes de la transformée de Fourier.

Annexe B

Eléments de quantification

On aborde ici le problème de la quantification d'une variable réelle ou vectorielle. Autrement dit, on cherche à *coder* un jeu de données sur un nombre fixé B de bits.

B.1 Généralités

B.1.1 Quantificateur scalaire

Soit $\mathsf{X} \subset \mathbb{R}$ un intervalle réel (éventuellement l'ensemble $\mathbb{R}$ tout entier). Soit $N \geq 2$ un entier. Un quantificateur à N niveaux est défini par la donnée de N éléments distincts $(\xi_1, \ldots, \xi_N) \in \mathsf{X}^N$ et par une fonction :

$$q \ : \ \mathsf{X} \to \{\xi_1, \ldots, \xi_N\}$$

Autrement dit, à tout x, la fonction q associe un point parmi N. L'entier N est le nombre de niveaux de quantification. L'ensemble de points $\{\xi_1, \ldots, \xi_N\}$ est appelé le *dictionnaire* associé à q. En définissant $C_j = q^{-1}(\xi_j)$ pour tout $j = 1, \ldots, N$, on peut écrire q sous la forme $q(x) = \sum_{j=1}^{N} \xi_j \mathbb{1}_{C_j}(x)$ où $\mathbb{1}_A$ représente l'indicatrice d'un ensemble A. Ainsi :

$$q(x) = \xi_j \quad \text{pour tout } x \in C_j \ .$$

Les ensembles C_j sont appelées des *cellules*. On en déduit qu'un quantificateur q est entièrement déterminé par la donnée d'un ensemble de points $\xi_1, \ldots, \xi_N$ et un ensemble de cellules $C_1, \ldots, C_N$ formant une partition de X. Lorsque X est un intervalle de la forme $\mathsf{X} = [a, b]$, un quantificateur est dit *uniforme* lorsque les cellules C_j sont des intervalles de longueurs identiques $(b - a)/N$.

B.1.2 Représentation binaire

Un *bit* est un élément de l'ensemble $\{0, 1\}$. Soit B un entier. Chaque nombre dans l'ensemble $\{0, 1, \ldots, 2^B - 1\}$ peut être représenté de manière unique par une suite de B bits correspondant à son écriture binaire. On pose dorénavant :

$$N = 2^B \ .$$

Il existe donc une fonction inversible $\mathcal{B} : \{\xi_1, \ldots, \xi_N\} \to \{0, 1\}^B$ qui à tout point ξ_j $(j = 1, \ldots, N)$ associe un vecteur de B bits, typiquement la représentation binaire de

$j-1$. Le nombre $B = \log_2 N$ représente donc le nombre de bits nécessaires pour stocker en mémoire la version quantifiée $q(x)$ d'une donnée x. Par exemple si $B = 3$ et $N = 2^3 = 8$,

$$\begin{aligned}
\mathcal{B} \ : \ \xi_1 &\mapsto (0,0,0) \\
\xi_2 &\mapsto (0,0,1) \\
\xi_3 &\mapsto (0,1,0) \\
\xi_4 &\mapsto (0,1,1) \\
\xi_5 &\mapsto (1,0,0) \\
\xi_6 &\mapsto (1,0,1) \\
\xi_7 &\mapsto (1,1,0) \\
\xi_8 &\mapsto (1,1,1) \ .
\end{aligned}$$

B.1.3 Mesure de distortion

L'objectif de la quantification est de fournir, à partir d'un nombre de bits B fixé, une description aussi fidèle que possible d'une entrée $x \in \mathsf{X}$ *a priori* inconnue. Une mesure naturelle de qualité est l'écart $|q(x) - x|$. Evidemment, le quantificateur ne peut être optimisé en fonction de x, sans quoi le problème serait trivial (choisir $C_1 = \{x\}$ et $\xi_1 = \{x\}$) et sans intérêt : le quantificateur n'aurait pas de raison d'être pertinent pour une nouvelle donnée x'. Une manière de contourner se problème est de supposer que l'entrée est une réalisation d'une variable *aléatoire* dont on connait la loi, et de chercher un quantificateur qui soit pertinent *en moyenne*.

Soit X une variable aléatoire réelle. On suppose dans ce qui suit que X admet une densité f dont le support est inclus dans X :

$$\mathbb{P}(X \in A) = \int_A f(x)dx \ , \ \forall A \ .$$

On suppose par ailleurs que $\mathbb{E}(X^2) < \infty$. On désigne par $\hat{X} = q(X)$ la version quantifiée de X. Pour tout quantificateur q, on définit la mesure de distortion suivante :

$$D(q) = \mathbb{E}\left[(\hat{X} - X)^2\right] \ .$$

Cette valeur est aussi appelée l'*erreur quadratique moyenne (EQM) de reconstruction*. Remarquons que d'autres mesures de distortions seraient également possibles, par exemple $\mathbb{E}\left[|\hat{X} - X|\right]$ ou encore $\mathbb{E}\left[(\hat{X} - X)^4\right]$. Néanmoins, l'erreur quadratique moyenne est de très loin la métrique la plus répandue, en particulier parce qu'elle permet d'obtenir quantité de résultats théoriques très utiles, comme nous le verrons plus loin.

B.1.4 Cellules de Voronoï

Pour tout N-uplet $(\xi_1, \ldots, \xi_N) \in \mathsf{X}^N$ et pour tout $j = 1, \ldots, N$, on définit la cellule de Voronoï associée à ξ_j par :

$$V_j = \{x \in \mathsf{X} \ : \ \forall k \neq j, \ |x - \xi_j| < |x - \xi_k|\} \ . \tag{B.1}$$

Il est facile de voir que V_j n'est rien d'autre que l'intervalle

$$V_j = \left] \frac{\xi_{j-1} + \xi_j}{2} \ , \ \frac{\xi_j + \xi_{j+1}}{2} \right[$$

pour tout $j = 2, \ldots, N-1$. Le quantificateur associé

$$q(x) = \sum_{j=1}^{N} \xi_j \mathbb{1}_{V_j}(x) \ .$$

est appelé quantificateur de Voronoï. Notons qu'en toute rigueur, les cellules $C_1, \ldots, C_N$ ne forment pas une partition de X car les bords de ces intervalles ne sont inclus dans aucune cellule. Néanmoins, comme X a une probabilité nulle d'être égal à l'un de ces points, cela est sans importance. Le lecteur scrupuleux pourra modifier la définition (B.1) afin d'inclure les bords des cellules dans l'une ou l'autre des cellules, de manière à former une partition. Il est clair que la manière de rattacher ces points à l'une ou l'autre des cellules n'a aucune influence sur l'EQM $D(q)$ associée au quantificateur obtenu.

La propriété ci-dessous montre que si les points du quantificateur sont fixés, alors le meilleur choix de possible de cellules est la partition de Voronoï.

Propriété 1. Soit $(\xi_1, \ldots, \xi_N) \in \mathsf{X}^N$. Soit q le quantificateur de Voronoï associé à $(\xi_1, \ldots, \xi_N)$. Alors,

$$D(q) \leq D(q')$$

pour tout quantificateur q' tel que $q'(\mathsf{X}) = \{\xi_1, \ldots, \xi_N\}$.

Démonstration. Remarquons que $|q(x) - x| = \min\{|\xi_\ell - x| \ : \ \ell = 1, \ldots, N\}$. Ainsi, pour tout k, $|\xi_k - x| \leq |q(x) - x|$. En particulier, $|q'(x) - x| \leq |q(x) - x|$. En élevant au carré et en intégrant cette relation, on conclut que $\mathbb{E}[(q'(X) - X)^2] \leq \mathbb{E}[(q(X) - X)^2]$. □

Exemple. Rappelons que le quantificateur uniforme sur l'intervalle $[0, 1]$ a pour $j$ème cellule l'intervalle $](j-1)/N, j/N[$. Ce quantificateur est donc de Voronoï lorsque ses points sont choisis pour que :

$$\xi_j = \frac{2j-1}{2N}$$

pour tout $j = 1, \ldots, N$.

B.2 Analyse des quantificateurs haute-résolution

B.2.1 L'intégrale de Bennett

L'EQM $D(q)$ ne possède malheureusement pas d'expression simple. Ainsi, il est généralement impossible de fournir une expression compacte et utilisable de la distortion D en fonction de la densité f, et encore moins de fournir une expression exacte des points ξ_j minimisant la distortion. Pour contourner cette difficulté, nous étudions le comportement asymptotique de l'EQM lorsque le nombre de niveaux de quantification tend vers l'infini, autrement dit lorsque le diamètre des cellules tend vers zéro. On parle de quantification *haute résolution*. La propriété 1 permet, sans restriction, de limiter l'analyse aux quantificateurs de Voronoï.

Considérons une suite de quantificateurs de Voronoï $(q_N)_{N\geq 2}$. Pour un N fixé, le quantificateur q_N est entièrement défini par la donnée de son dictionnaire. Ce dictionnaire est une fonction de N, nous le noterons donc $\{\xi_{1,N}, \ldots, \xi_{N,N}\}$. Pour tout ensemble A, la quantité

$$\mu_N(A) = \frac{1}{N} \text{ card } \{j = 1, \ldots, N \ : \ \xi_{j,N} \in A\}$$

représente le nombre de points contenus dans A, normalisé par le nombre total de points. Il est facile de voir que μ_N est une mesure de probabilité sur l'ensemble des boréliens. Supposons qu'il existe une certaine densité ζ sur $\mathbb{R}$ telle que pour tout intervalle $A = [a, b]$,

$$\lim_{N\to\infty} \mu_N(A) = \int_A \zeta(x)dx \ .$$

En langage probabiliste, on dit que μ_N *converge faiblement* vers la mesure $\zeta(x)dx$. De manière un peu moins formelle, cela revient à supposer que lorsque N est grand,

$$\text{Nombre de points dans un petit intervalle } [x, x+\delta] \ \simeq N\,\zeta(x)\,\delta \ .$$

Nous désignerons ζ comme la *densité asymptotique de points* de quantification. Sous certaines conditions, nous avons le résultat suivant dû à Bennett (Bennett, 1948) :

$$\lim_{N\to\infty} N^2 D(q_N) = I(\zeta) \ , \tag{B.2}$$

où $I(\zeta)$ est donnée par :

$$I(\zeta) = \frac{1}{12} \int \frac{f(x)}{\zeta(x)^2} dx \ . \tag{B.3}$$

L'intégrale ci-dessus est appelée l'*intégrale de Bennett*. L'équation (B.2) implique tout d'abord que l'erreur de reconstruction tend vers zéro lorsque $N \to \infty$, ce qui est évidemment un résultat attendu. En outre, la vitesse de convergence est en $1/N^2$. Enfin, elle fournit une expression asymptotique très explicite qui permet de faire le lien entre la loi de la variable X et la densité asymptotique de points.

Le résultat de Bennett est vrai sous des hypothèses assez générale et l'on pourra se reporter à l'article (Na & Neuhoff, 1995) ou à l'ouvrage (Graf & Luschgy, 2000) pour une preuve détaillée. Dans ce cours, nous nous contentons de démontrer (B.2) sous des hypothèses plus restrictives, en particulier, en supposant que X est borné. Cette hypothèse pourrait en réalité être relaxée.

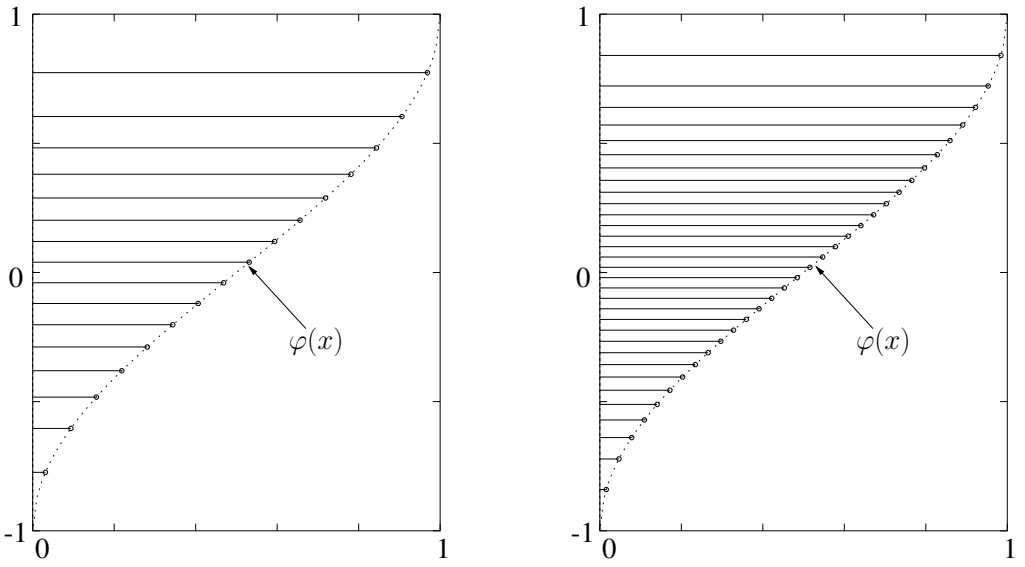
Hypothèse 1. a) X est un intervalle borné.

b) La densité de probabilité f est lipschitzienne.

c) Il existe une fonction $\varphi : [0, 1] \to \mathbb{R}$ strictement croissante et de classe C^2 telle que pour tout N,

$$\forall j = 1 \ldots, N, \quad \xi_{j,N} = \varphi\left(\frac{2j-1}{2N}\right) \ .$$

Notons qu'on peut toujours exprimer les points d'un quantificateur comme une fonction φ des points $(2j-1)/(2N)$ du quantificateur uniforme, cela n'est aucunement restrictif. L'hypothèse 1.c) inmplique de surcroît que cette fonction est *la même* quel que soit N. Par cette hypothèse, on introduit une cohérence dans la famille de quantificateur $(q_N)_N$ comme le représente la figure ci-dessous.



(a) N=16 (b) N=32

FIGURE B.1 – Exemple de famille de quantificateurs sur $\mathsf{X} = [-1, 1]$ vérifiant l'hypothèse 1.c). On se donne une fonction φ sur $[0, 1]$ qu'on échantillonne aux points $(2j-1)/(2N)$. Les points du quantificateurs sont les ordonnées $\varphi((2j-1)/(2N))$. On voit que la densité de points est d'autant plus importante que la dérivée de la fonction φ est faible.

Théorème 1. Soit $(q_N)_{N\geq 2}$ une famille de quantificateurs de Voronoï. Supposons l'hypothèse 1 satisfaite. Alors la famille $(q_N)_{N\geq 2}$ admet pour densité asymptotique de points la fonction $\zeta = (\varphi^{-1})'$. De plus, l'équation (B.2) est vraie.

Démonstration. • Commençons par montrer que la densité asymptotique de points est égale à $(\varphi^{-1})'$. Soit $A = [a, b]$ un intervalle quelconque. Par définition, $N\mu_N(A)$ est le nombre de point $\xi_{j,N}$ tels que $a \leq \xi_{j,N} \leq b$. Par hypothèse, cet encadrement revient à $a \leq \varphi(j/N - 1/(2N)) \leq b$ soit par croissance de φ, $N\varphi^{-1}(a) + \frac{1}{2} \leq j \leq N\varphi^{-1}(b) + \frac{1}{2}$. Ainsi,

$$N\varphi^{-1}(b) - N\varphi^{-1}(a) - 1 \leq N\mu_N(A) \leq N\varphi^{-1}(b) - N\varphi^{-1}(a) + 1 \ .$$

Quand $N \to \infty$, $\mu_N(A)$ converge vers $\varphi^{-1}(b) - \varphi^{-1}(a) = \int_A (\varphi^{-1})'$. Ceci prouve le premier point.

• Démontrons maintenant la formule de Bennett.

$$D(q_N) = \int (x - q_N(x))^2 f(x)dx = \sum_{j=1}^{N} \int_{V_{j,N}} (x - \xi_{j,N})^2 f(x)dx \ , \tag{B.4}$$

où $V_{j,N}$ représente la $j$ème cellule de Voronoï. Pour tout $j = 2, \ldots, N-1$, on a d'après (B.1) : $V_{j,N} =]a_{j-1,N}, a_{j,N}[$ où $a_{j,N} = (\xi_{j,N} + \xi_{j+1,N})/2$. Par conséquent,

$$\begin{aligned}
\int_{V_{j,N}} (x - \xi_{j,N})^2 f(x)dx &= f(\xi_{j,N}) \int_{V_{j,N}} (x - \xi_{j,N})^2 dx + r_{j,N} \\
&= \frac{f(\xi_{j,N})}{3}[(a_{j,N} - \xi_{j,N})^3 - (a_{j-1,N} - \xi_{j,N})^3]dx + r_{j,N} \\
&= \frac{f(\xi_{j,N})}{24}[(\xi_{j+1,N} - \xi_{j,N})^3 + (\xi_{j,N} - \xi_{j-1,N})^3] + r_{j,N} \ ,
\end{aligned}$$

où $r_{j,N} = \int_{V_{j,N}} (x - \xi_{j,N})^2 (f(x) - f(\xi_{j,N}))dx$ est un reste qui vérifie, si C est la constante de Lipschitz de f,

$$|r_{j,N}| \leq C \int_{V_{j,N}} |x - \xi_{j,N}|^3 dx \leq C \left(\frac{\xi_{j+1,N} - \xi_{j,N}}{2} \right)^4 \leq C(\sup_{t \in [0,1]} \varphi'(t))^4 N^{-4} \ .$$

L'inégalité ci-dessus montre que les restes $r_{j,N}$ ne jouent asymptotiquement aucun rôle, en ce sens que $\sum_j r_{j,N} = O(N^{-3})$. Par le même type d'argument, on peut facilement montrer que les premier ($j = 1$) et dernier ($j = N$) termes de la somme (B.4) sont asymptotiquement négligeables, si bien que :

$$N^2 D(q_N) = \frac{N^2}{24} \sum_{j=2}^{N-1} f(\xi_{j,N})[(\xi_{j+1,N} - \xi_{j,N})^3 + (\xi_{j,N} - \xi_{j-1,N})^3] + o(1)$$

où $o(1)$ désigne un terme qui tend vers zéro lorsque $N \to \infty$. En séparant la somme ci dessus en deux sommes, puis en réindexant l'une d'entre elle, on obtient :

$$N^2 D(q_N) = \frac{N^2}{12} \sum_{j=1}^{N-1} f(\xi_{j,N})(\xi_{j+1,N} - \xi_{j,N})^3 + o(1)$$

où l'on a inclus les "bords" de la somme dans le $o(1)$. Par le biais d'un développement de Taylor autour de j/N,

$$\xi_{j,N} = \varphi \left(\frac{j}{N} - \frac{1}{2N} \right) = \varphi(\frac{j}{N}) + \varphi'(\frac{j}{N})/(2N) + \epsilon_{j,N}$$

où $|\epsilon_{j,N}| \leq \sup |\varphi''| N^{-2}$ est le reste du développement de Taylor, qui s'avère négligeable. En développant de même $\xi_{j+1,N}$, on trouve que la différence $\xi_{j+1,N} - \xi_{j,N}$ coïncide avec $\varphi'(\frac{j}{N})$ plus un reste négligeable. Ainsi,

$$N^2 D(q_N) = \frac{1}{12\,N} \sum_{j=1}^{N-1} f(\varphi(j/N))(\varphi'(j/N))^3 + o(1) \ .$$

La somme de Riemann ci-dessus converge vers l'intégrale $\int_0^1 f(\varphi(t))(\varphi'(t))^3 dt$. Par le changement de variable $x = \varphi(t)$, cette intégrale est égale à $\int f(x)(\varphi'(\varphi^{-1}(x)))^2 dt$ et coïncide bien avec (B.3). □

B.2.2 Quantificateur asymptotiquement optimal

Proposition B.1. *L'intégrale de Bennett $I(\zeta)$ satisfait $I(\zeta) \geq I^\star$ où*

$$I^\star = \frac{1}{12} \left(\int f^{1/3} \right)^3 \ .$$

De plus, la borne $I^\star$ est atteinte lorsque la densité asymptotique de points coïncide avec la fonction :

$$\zeta^*(x) = \frac{f(x)^{1/3}}{\int f(y)^{1/3} dy} \ .$$

Démonstration. L'inégalité de Hölder implique que

$$\int \left(\frac{f}{\zeta^2}\right)^{1/3} \zeta^{2/3} \leq \left(\int \frac{f}{\zeta^2}\right)^{1/3} \left(\int \zeta\right)^{2/3} .$$

Puisque ζ est une densité, $\int \zeta = 1$. Le membre de droite de l'inégalité ci-dessus est égal à $\left(\int \frac{f}{\zeta^2}\right)^{1/3}$. Le membre de gauche est égal à $\int f^{1/3}$. Ainsi, $\int \frac{f}{\zeta^2} \geq \left(\int f^{1/3}\right)^3$ ce qui conduit à la borne de l'énoncé. On voit immédiatement que la densité ζ^* permet d'atteindre cette borne. □

Nous faisons les remarques suivantes :

— *Effet d'un changement d'échelle.* Posons $Y = aX + b$ où $a \neq 0$ et b sont deux scalaires. La densité de probabilitéf_Y de Y est liée à la densité f_X de X par la formule $f_Y(y) = a^{-1} f_X((y-b)/a)$. Si l'on note $I_Y^\star$ et $I_X^\star$ les distortions asymptotiques minimales associées à Y et X respectivement, un changement de variable montre que $I_Y^\star = a^2 I_X^\star$. En particulier, il est clair que la distortion asymptotique minimale ne dépend pas de l'espérance $E(X)$. Dans la suite, les variables seront toujours supposées centrées $E(X) = 0$.

— *Cas des variables gaussiennes.* Soit $X \sim \mathcal{N}(0, \sigma^2)$ une variable gaussienne de moyenne nulle et de variance $\sigma^2 > 0$. La variable $\sigma^{-1}X$ suit une loi $\mathcal{N}(0,1)$. D'après ce qui a été dit plus haut, la distortion asymptotique minimale $I^\star$ associée à X s'écrit $I^\star = \sigma^2 I^\star_{\mathcal{N}(0,1)}$ où $I^\star_{\mathcal{N}(0,1)}$ est la distortion asymptotique minimale de la loi $\mathcal{N}(0,1)$. On peut aisément calculer cette constante en remarquant que :

$$\int \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right)^{1/3} dx = \frac{1}{\sqrt{2\pi}^{1/3}} \int e^{-\frac{x^2}{2\times 3}} dx = \frac{\sqrt{2\pi 3}}{\sqrt{2\pi}^{1/3}} \int \frac{1}{\sqrt{2\pi 3}} e^{-\frac{x^2}{2\times 3}} dx = (2\pi)^{1/3}\sqrt{3} .$$

Nous obtenons $I^\star_{\mathcal{N}(0,1)} = \frac{1}{12}\{(2\pi)^{1/3}\sqrt{3}\}^3 = \sqrt{3}\pi/2$. Finalement,

$$I^\star = \frac{\sqrt{3}\pi}{2}\, \sigma^2 . \tag{B.5}$$

On remarque par ailleurs que la densité optimale de point $\zeta^\star$ permettant de quantifier une gaussienne de manière (asymptotiquement) optimale n'est rien d'autre qu'une gaussienne, de variance $3\sigma^2$:

$$\zeta^\star(x) = \frac{1}{\sqrt{6\pi}} e^{-x^2/6\sigma^2} . \tag{B.6}$$

Il s'agit d'une règle précieuse pour construire des quantificateurs ayant de bonnes performances en pratique.

B.3 Quantification d'un vecteur de données

On se demande désormais comment quantifier un vecteur de données

$$\boldsymbol{X} = (X_1, \ldots, X_n)^T \in \mathsf{X}^n .$$

Une approche générale consisterait à chercher des quantificateurs vectoriels, c'est à dire une fonction de X^n dans un sous-ensemble fini de X^n, sans s'imposer *a priori* de restriction particulière sur le choix de cette fonction. Dans ce chapitre, nous nous limiterons à des structures de quantificateurs construits à partir de n quantificateurs scalaires. Ce type de structure est communément utilisé dans les schémas de compression usuels (citons par exemple le format JPEG) pour sa simplicité et ses bonnes performances pratiques.

Une vaste littérature existe sur la quantification vectorielle dans un cadre plus général, mais elle dépasse largement le contexte de ce cours. On pourra se reporter à l'ouvrage (Gersho & Gray, *Vector quantization and signal compression*, 1992) pour plus de détails.

B.3.1 Allocation optimale des bits

Dorénavant, nous ferons l'hypothèse que $\boldsymbol{X}$ est un **vecteur gaussien** centré.

On suppose que l'on dispose de B bits pour coder le vecteur $\boldsymbol{X}$. La structure la plus naturelle consiste à quantifier le vecteur $\boldsymbol{X}$ composante par composante. On se donne n quantificateurs $q^{(1)}, \ldots, q^{(n)}$ et on note $\hat{X}_i = q^{(i)}(X_i)$ la version quantifiée de X_i par le $i$ème quantificateur, comme l'indique la figure B.2. Un tel quantificateur vectoriel est appelé un quantificateur-produit. On note $\hat{\boldsymbol{X}}$ le vecteur dont la $i$ème composante est $\hat{X}_i$. On désigne

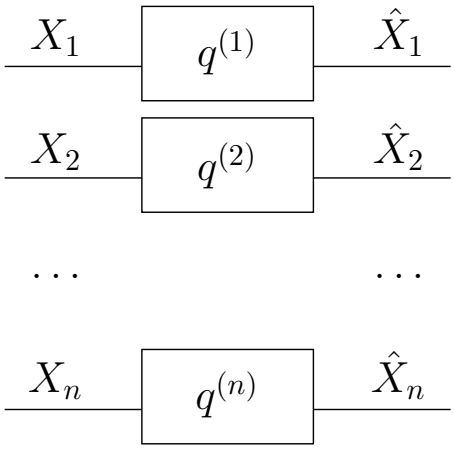


FIGURE B.2 – Structure d'un quantificateur-produit.

par b_i le nombre de bits du $i$ème quantificateur. Le nombre de niveaux correspondant est égal à $N_i = 2^{b_i}$. Puisque le nombre de bits disponibles est égal à B, le choix des b_i est sujet à la contrainte :

$$\sum_{i=1}^{n} b_i \leq B \ . \tag{B.7}$$

On se pose le problème d'optimisation suivant, appelé le problème de l'allocation des bits. Sous la contrainte (B.7), trouver les valeurs de $b_1, \ldots, b_n$ et les quantificateurs $q^{(1)}, \ldots, q^{(n)}$ minimisant l'erreur quadratique moyenne de reconstruction :

$$\mathbb{E}\left(\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|^2\right) = \sum_{i=1}^{n} \mathbb{E}[(\hat{X}_i - X_i)^2] \ .$$

Posé de cette manière, le problème ne peut avoir de solution tractable car encore une fois, le critère ci-dessus n'admet pas d'expression simple. Toutefois, lorsque N_i est suffisamment grand et lorsque la densité des points de quantification approche la densité optimale (B.6), l'équation (B.5) implique que :

$$N_i^2 \, \mathbb{E}[(\hat{X}_i - X_i)^2] \simeq \frac{\sqrt{3}\pi}{2} \, \sigma_i^2 \ ,$$

où σ_i^2 est la variance de X_i. En utilisant le fait que $N_i = 2^{b_i}$, l'approximation ci-dessus s'écrit de manière équivalente :

$$\mathbb{E}\left(\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|^2 \right) \simeq \sum_{i=1}^{n} \frac{\sqrt{3}\pi}{2} \, \sigma_i^2 \, 2^{-2b_i} \ .$$

On choisit donc de reformuler le problème de l'allocation des bits sous la forme :

$$\text{Minimiser } \frac{\sqrt{3}\pi}{2} \sum_{i=1}^{n} \sigma_i^2 \, 2^{-2b_i} \text{ sous contrainte } \sum_i b_i \leq B \ . \tag{B.8}$$

En toute rigueur, la solution du problème est à rechercher dans $\mathbb{N}^n$, c'est à dire avec la contrainte que b_i est entier pour tout i. Une telle contrainte est toutefois délicate à prendre en compte et ne permet pas d'obtenir une forme évidente de la solution. On cherchera donc la solution de ce problème sans s'imposer que b_i est entier. Naturellement, les b_i trouvés devront être arrondis si l'on souhaite une solution pratique. Par exemple, si la solution théorique indique $b_i = 0,2$bits, on choisira simplement $b_i = 0$, c'est à dire que l'on éliminera simplement la composante X_i du codage. Notons que lorque B est grand, l'arrondi a très peu d'effet sur la valeur finale de la distortion.

Proposition B.2. *Soit $\rho^2 = (\prod_{i=1}^n \sigma_i^2)^{1/n} > 0$ la moyenne géométrique des variances et $\bar{b} = B/n$ le nombre moyen de bits disponibles par composante. Le minimum global du problème (B.8) est égal à*

$$\mathcal{D}_{\boldsymbol{X}} = \frac{\sqrt{3}\pi n}{2} \rho^2 2^{-2\bar{b}}$$

et est atteint pour l'allocation définie par :

$$b_i = \bar{b} + \frac{1}{2} \log_2 \frac{\sigma_i^2}{\rho^2} \ , \ \forall i \ .$$

Autrement dit, le nombre de niveaux N_i est proportionnel à l'écart-type σ_i : $N_i \propto \sigma_i$.

Démonstration. La preuve repose sur l'inégalité des moyennes arithmétiques et géométriques :

$$\frac{1}{n} \sum_{i=1}^{n} a_i \geq (\prod_{i=1}^{n} a_i)^{1/n} \ . \tag{B.9}$$

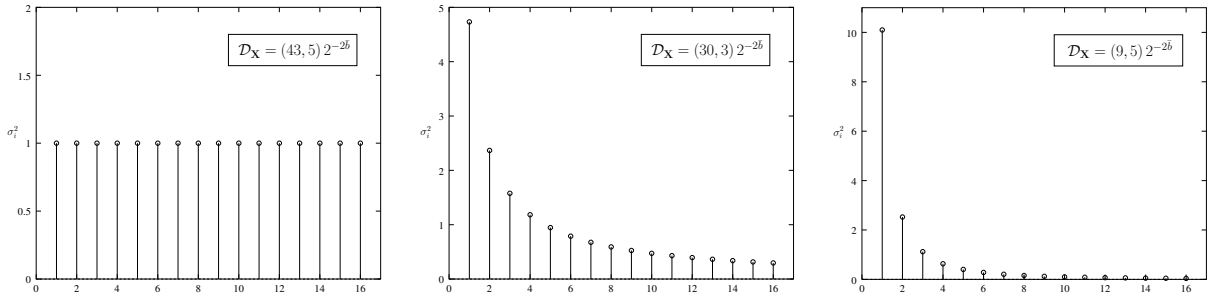
Cette inégalité conduit à :

$$\sum_{i=1}^{n} \sigma_i^2 \, 2^{-2b_i} \geq n \left(\prod_{i=1}^{n} \sigma_i^2 \, 2^{-2b_i} \right)^{1/n} = n\rho^2 2^{-2\sum_i b_i/n} \geq n\rho^2 2^{-2\bar{b}}$$

ce qui prouve l'expression du minimum. Il reste à montrer que ce minimum est atteint pour l'allocation fournie dans l'énoncé de la proposition. La vérification est immédiate et laissée au lecteur. □

Remarque : Soit $\boldsymbol{X}$ un vecteur aléatoire d'*énergie* fixée $E = \mathbb{E}(\|\boldsymbol{X}\|^2)$. L'inégalité (B.9) implique que $\rho^2 \leq E$ avec égalité lorsque toutes les variances sont égales : $\sigma_i^2 = E/n$. Ainsi pour une énergie E fixée, le **pire cas** en termes de qualité de compression est le cas de variances égales. Cette remarque donne l'intuition suivante, importante pour la suite. Afin de compresser efficacement un vecteur, *il est intéressant de lui appliquer préalablement une transformation qui concentre son énergie sur un nombre de composantes aussi petit que possible*.

La figure B.3.1 illustre ce phénomène.



(a) (b) (c)

FIGURE B.3 – Evaluation de la distortion asymptotique minimale pour $N = 16$ trois répartitions différentes des variances $\sigma_1^2, \ldots, \sigma_{16}^2$ vérifiant toutes $\sum_i \sigma_i^2 = 16$. (a) variances uniformes - (b) variances σ_i^2 décroissantes en $1/i$ - (c) variances σ_i^2 décroissantes en $1/i^2$. La distortion décroît d'autant plus vite que l'énergie est concentrée sur peu de composantes.

B.3.2 Quantification par transformée

Le principe du codage par transformée est illustré par la figure B.4. L'idée sous-jacente

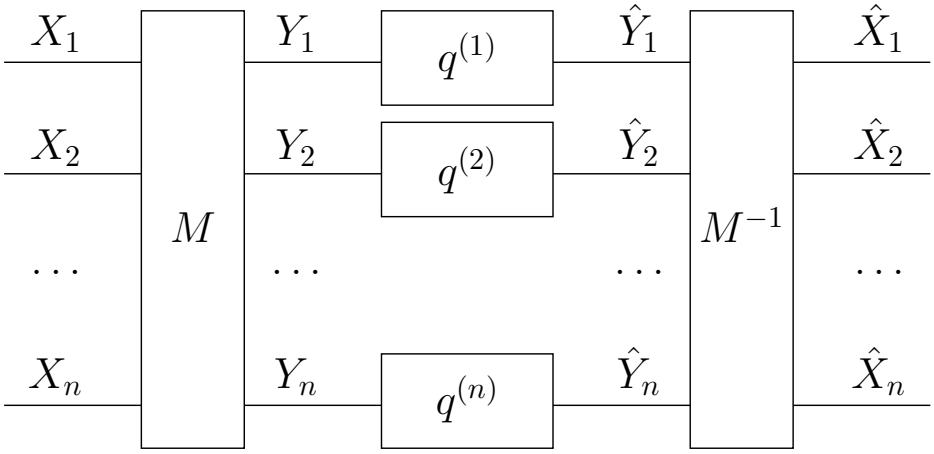


FIGURE B.4 – Structure d'un quanticateur par transformée.

est que l'on peut réduire la distortion induite par l'étape de quantification vue au paragraphe précédent si, au lieu de quantifier le vecteur $\boldsymbol{X}$ lui-même, on quantifie une version transformée $\boldsymbol{Y} = M\boldsymbol{X}$, où M est une matrice que nous supposerons orthogonale :

$$M^{-1} = M^T \ .$$

Nous notons $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^T$ le vecteur quantifié des entrées transformées : $\hat{Y}_i = q^{(i)}(Y_i)$ où $q^{(i)}$ est un quantificateur sur b_i bits et $\sum_i b_i \leq B$. La reconstitution se fait par application la transformée inverse M^{-1} aux points de quantification : $\hat{\boldsymbol{X}} = M^{-1}\hat{\boldsymbol{Y}}$. Les questions qui se posent sont les suivantes : quelle transformation M permet de minimiser l'EQM de reconstruction, et quelle est l'amélioration obtenue par rapport à une simple quantification composante par composante ? D'après ce qui a été dit plus haut, nous savons déjà qu'une "bonne" transformation M est une transformation qui concentre un maximum d'énergie sur un minimum de composantes. Essayons maintenant d'être plus précis.

Comme précédemment, nous faisons l'hypothèse que $\boldsymbol{X}$ est un vecteur gaussien centré, et nous notons Σ sa matrice de covariance :

$$\boldsymbol{X} \sim \mathcal{N}(0, \Sigma) \ . \tag{B.10}$$

Rappelons qu'une transformation linéaire d'un vecteur gaussien reste un vecteur gaussien (*cf.* cours de probabilités). Le vecteur $\boldsymbol{Y} = M\boldsymbol{X}$ est donc un vecteur gaussien et sa matrice de covariance est donnée par $\tilde{\Sigma} = M\Sigma M^T$:

$$\boldsymbol{Y} \sim \mathcal{N}(0, \tilde{\Sigma}) \ \text{ où } \tilde{\Sigma} = M\Sigma M^T \ .$$

Nous désignons par $\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_n^2$ les éléments diagonaux de $\tilde{\Sigma}$. En particulier, nous avons par définition de la matrice de covariance :

$$Y_i \sim \mathcal{N}(0, \tilde{\sigma}_i^2) \ .$$

On s'intéresse maintenant à la réduction de l'EQM de reconstruction $\mathbb{E}\left(\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|^2\right)$. Puisqu'une transformation orthogonale de modifie pas la norme,

$$\mathbb{E}\left(\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|^2\right) = \mathbb{E}\left(\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}\|^2\right) \ .$$

Ainsi, il suffit de choisir une matrice M pour laquelle l'EQM de reconstruction de $\boldsymbol{Y}$ est petite. Encore une fois, nous remplaçons cette EQM par son expression asymptotique. Nous avons vu au paragraphe précédent que pour une allocation bien choisie des bits associés à chaque quantificateur, nous avons $\mathbb{E}\left(\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}\|^2\right) \simeq \mathcal{D}_{\boldsymbol{Y}}$ où

$$\mathcal{D}_{\boldsymbol{Y}} = \frac{\sqrt{3}\pi n}{2}(\prod_{i=1}^{n} \tilde{\sigma}_i^2)^{1/n} 2^{-2\bar{b}} \ .$$

Par conséquent, on se pose le problème d'optimisation suivant.

$$\text{Minimiser } \mathcal{D}_{\boldsymbol{Y}} \text{ sous la contrainte que } M \text{ est orthogonale.} \tag{B.11}$$

Rappelons qu'une matrice de covariance est symétrique semi-définie positive ($x^T\Sigma x \geq 0$ pour tout x) et donc diagonalisable dans une base orthonormée. De manière équivalente, on peut écrire qu'il existe une matrice orthogonale P telle que :

$$\Sigma = P\Lambda P^T$$

où Λ est une matrice diagonale contenant les valeurs propres de Σ.

Proposition B.3. *Le minimum global du problème (B.11) est égal à* $\frac{\sqrt{3}\pi n}{2}(\det\Sigma)^{1/n}\,2^{-2\bar{b}}$. *Ce minimum est atteint lorsque* $M = P^T$.

Démonstration. La preuve est une conséquence de l'inégalité d'Hadamard, qui établit que le déterminant d'une matrice semi-définie positive est plus petit que le produit des éléments diagonaux. Cela se lit :

$$\det A \leq \prod_{i=1}^{n} a_{i,i} \tag{B.12}$$

où $A = [a_{i,j}]$ est une matrice semi-définie positive $n \times n$. Cette inégalité est prouvée plus bas. A partir de (B.12), on montre que :

$$\mathcal{D}_{\boldsymbol{Y}} \geq \frac{\sqrt{3}\pi n}{2}(\det\tilde{\Sigma})^{1/n}\,2^{-2\bar{b}} = \frac{\sqrt{3}\pi n}{2}(\det\Sigma)^{1/n}\,2^{-2\bar{b}}$$

où l'on a utilisé le fait que $\det\tilde{\Sigma} = \det P\Sigma P^{-T} = \det\Sigma$. Il reste à montrer que la borne est atteinte lorsque $M = P^T$. Dans ce dernier cas, $\tilde{\Sigma} = M\Sigma M^T = P^T\Sigma P = \Lambda$. La matrice $\tilde{\Sigma}$ est donc diagonale. Le produit $\prod_{i=1}^{n}\tilde{\sigma}_i^2$ de ses éléments diagonaux coincide avec son déterminant $\det\tilde{\Sigma} = \det\Sigma$. Ainsi, la borne est atteinte.

Preuve de l'inégalité d'Hadamard : Commençons par le cas simple où $a_{i,i} = 1$ pour tout i. Appelons $\lambda_1, \ldots, \lambda_n$ les valeurs propres de A. L'inégalité (B.9) implique que :

$$\det A = \prod_{i=1}^{n}\lambda_i \leq \left(\frac{1}{n}\sum_{i=1}^{n}\lambda_i\right)^n = \left(\frac{1}{n}\mathrm{trace}(A)\right)^n = 1\ ,$$

ce qui prouve l'inégalité dans ce cas particulier. Traitons maintenant le cas général. On suppose que A est définie positive (dans le cas contraire, le déterminant est nul est l'inégalité est triviale). Dans ce cas, $a_{i,i} \neq 0$ pour tout i. Soit D la matrice diagonale dont le $i$ème coefficient vaut $a_{i,i}^{-1/2}$. D'après la preuve précédente, $\det DAD \leq 1$. Or $\det DAD = \det A(\det D)^2 = \det A\left(\prod_i^n a_{i,i}^{-1/2}\right)^2$. L'inégalité d'Hadamard est donc prouvée. □

Le résultat ci-dessus a l'implication suivante. Désignons par

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

les valeurs propres dans un ordre décroissant, et notons $e_1, \ldots, e_n$ les vecteurs propres respectivement associés à chacune de ces valeurs propres. On a $P = [e_1, \ldots, e_n]$. Ainsi, les entrées du vecteur $\boldsymbol{Y} = P^T\boldsymbol{X}$ ne sont rien d'autre que les coordonnées du vecteur $\boldsymbol{X}$ dans la base $e_1, \ldots, e_n$:

$$Y_i = \langle \boldsymbol{X}, \boldsymbol{e}_i\rangle\ .$$

La matrice de covariance de $\boldsymbol{Y}$ est égale à $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ et en particulier, la variance de Y_i est égale à λ_i. D'après la proposition B.2, le nombre optimal de points N_i du $i$ème quantificateur doit être proportionnel à $\sqrt{\lambda_i}$. La figure B.5 illustre cette structure optimale de quantification par transformée. De ce qui précède, on déduite les règles générales suivantes, qui nous serviront pour la construction de quantificateurs :

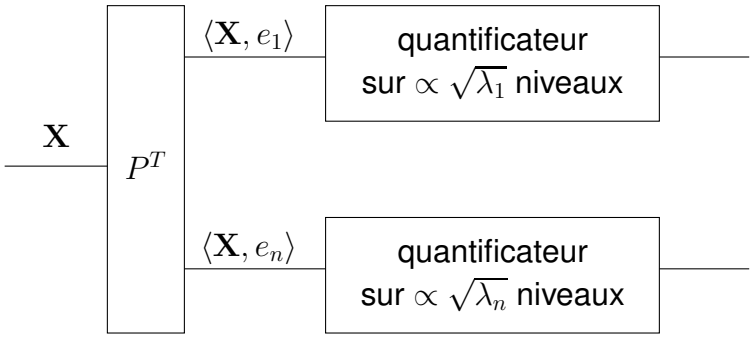


FIGURE B.5 – Quantificateur par transformée optimal

1. Appliquer un changement de base pertinent, adapté à la base des vecteurs propres de la matrice de covariance de l'entrée ;
2. Quantifier finement les coordonnées $\langle \boldsymbol{X}, \boldsymbol{e}_i \rangle$ qui correspondent aux plus grandes valeurs propres ; Quantifier plus grossièrement voire éliminer les coordonnées $\langle \boldsymbol{X}, \boldsymbol{e}_i \rangle$ qui correspondent aux plus faibles valeurs propres.