Signal Processing

E. Le Pennec

2014-2015

Introduction

Signal Processing:

- signal: sound, image, seismic trace, prices,...
- classical processing: transmission, denoising, coding,...
- other processing: signal analysis (detection, pattern recognition, segmentation...), synthesis,...

This course: Introduction to (classical) Signal Processing.

Mathematically correct but some (technical) details swept under the carpet.

Focus on selected real-life issues:

- **FFT:** Fast implementation of Discrete Fourier Transform, an ubiquitous algorithm in the discrete world (JPEG, MPEG,...).
  - Stress importance of the Fourier transform in (TI) Signal Processing,
  - Explain the discretization effects using (a simplified) Distribution Theory,
  - Analyze Discrete Filtering up to basic z transform.
- **Voice processing:** Vocoder (synthesis and compression)
  - Justify the time-frequency analysis (and representation) by the non-stationarity
  - Introduce stochastic (locally) stationary modeling
  - Describe the LPC modeling for synthesis and basic compression
- **Image compression:** GIF, PNG and JPEG (no need for justification...)
  - Lossless Image Coding: Shannon theory and predictive
  - Lossy Image Coding: quantization and transform coding
- Trends in Signal Processing:
  - Sparse modeling
  - Inverse problem
  - Segmentation

– Learning and features

Part I

FFT: Analog Signal Processing, Discretization and Digital Signal Processing

Chapter 1

Analog Signal, LTI and Fourier transform

1.1 LTI and Analog signal

System: input an *signal* in a space $\mathcal{I}$, process it and output another *signal* in a space $\mathcal{O}$ (Operator)

Examples: Heat equation, radio communication, electrical circuit, optical lens, CD, TV...

Very general framework: We focus here on the case where the *signals* are analog (think of continuous function) and the operators are Linear and Translation Invariant operators.

Linearity: $L(\lambda f + g) = \lambda L(f) + L(g)$.

Translation Invariance: Denote by $\tau_\Delta f(t) = f(t - \Delta)$ the Δ-shifted version of f, we assume that $L(\tau_\Delta f) = \tau_\Delta(Lf)$ (the image of a translated signal is the translation of the image of the original signal).

Continuity: Require to specify the input set $\mathcal{I}$, the output set $\prime$ as well as their topology... Much more complex that it may look like. All the interesting LTI operators we will see satisfy this continuity assumption in a generalized sense (distribution).

Such a LTI operator is called a filter, we will see later why.

Examples: Heat equation, radio transmission, equalizer...

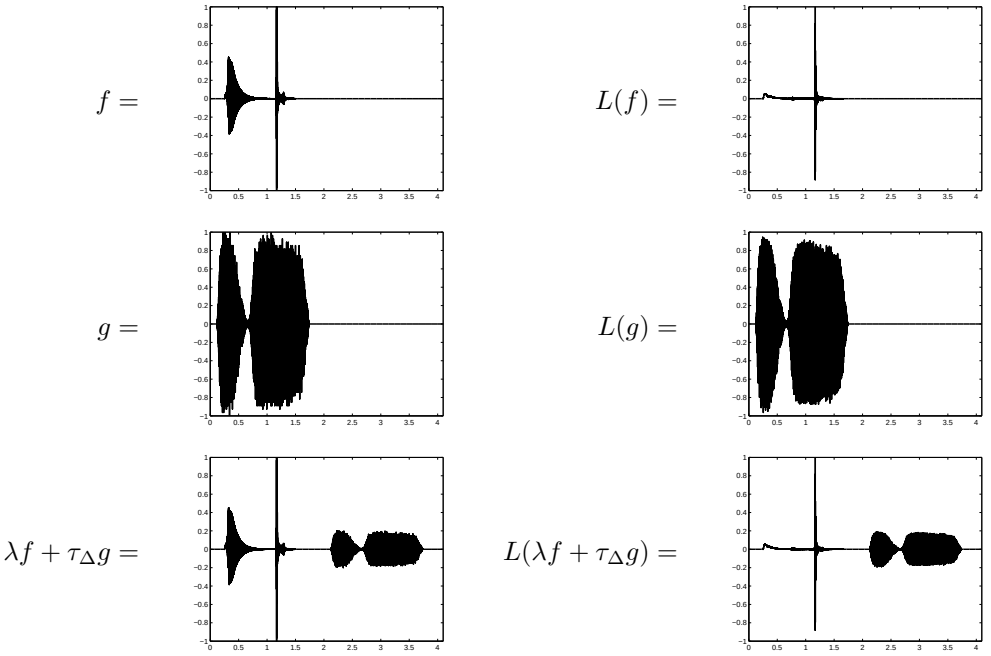


Figure 1.1: LTI properties

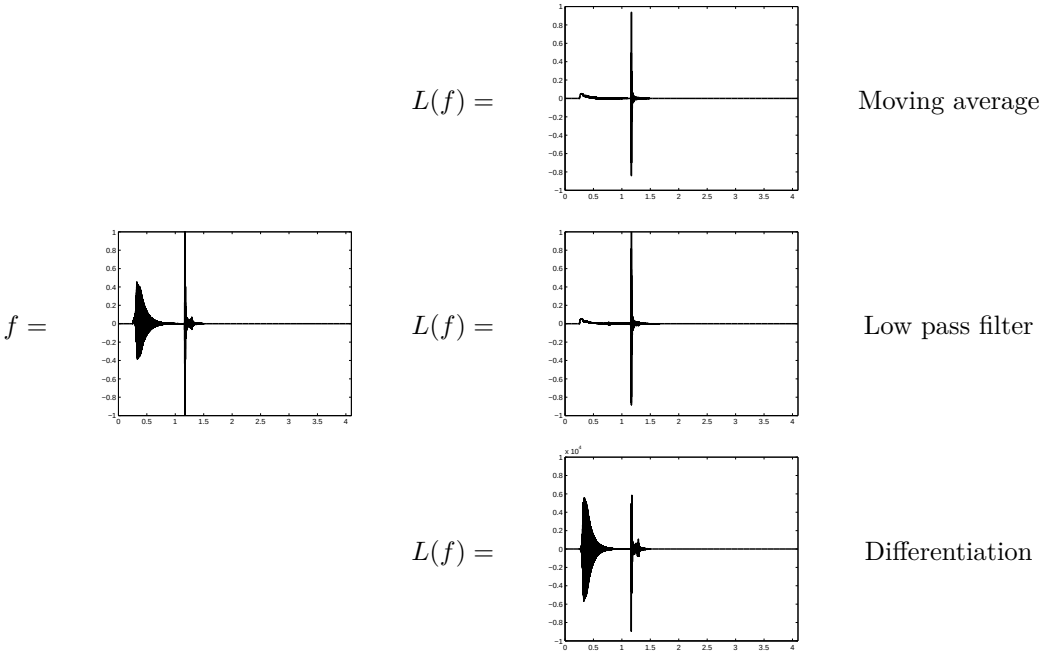


Figure 1.2: LTI examples

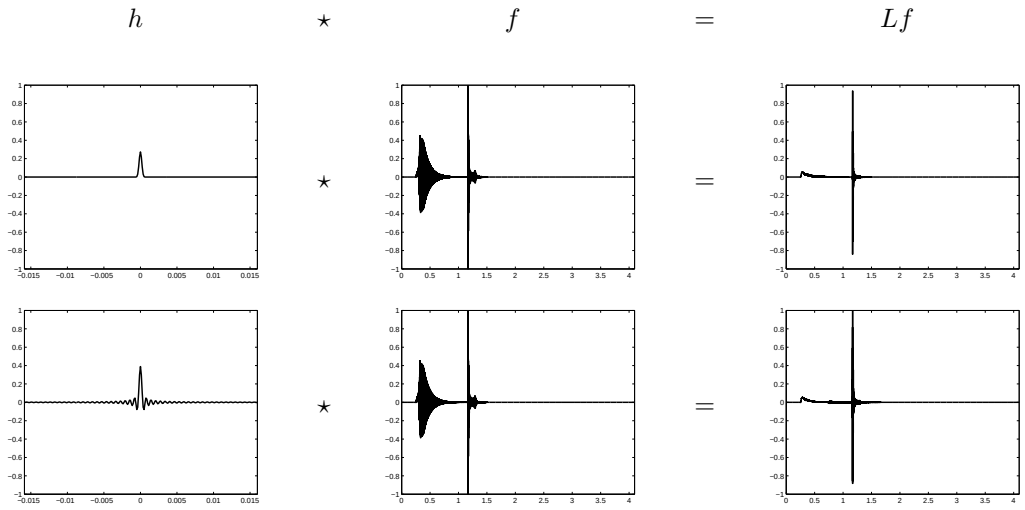


Figure 1.3: Impulse response

Fundamental example: convolution with a function h

$$Lf = h \star f \Leftrightarrow Lf(t) = \int f(u)h(t-u)\mathrm{d}u.$$

h is called the impulse response of the operator. This case is very generic but this definition supposed that the convolution is well defined...

Example: Local smoothing and low pass filter.

Proposition 1.1: a LTI operator commutes with the derivation operator.

Proof: $L(\frac{\tau_h f - f}{h}) = \frac{\tau_h Lf - Lf}{h}$

Stability of a filter: continuity for specific norms. Most classical one is BIBO (Bounded Input Bounded Output) which corresponds to the property

$$\|f\|_\infty < +\infty \implies \|Lf\|_\infty < +\infty$$

Proposition 1.2: If $Lf = h \star f$ then this is equivalent to $\|h\|_1 < +\infty$.

Causality: an important property for instance for a real-time process is that the output should not depend on the future, this property is called causality.

Proposition 1.3: If $Lf = h \star f$ then this equivalent to $h(u) = 0$ if $u < 0$.

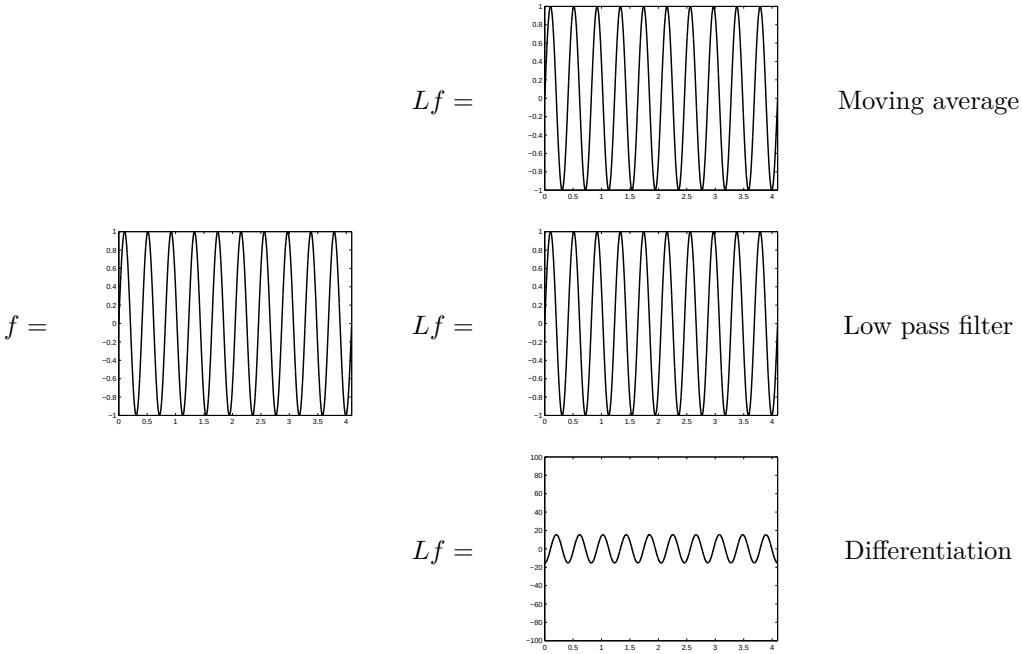


Figure 1.4: Exponential and LTI

Exponential input: Let $e_\sigma(t) = e^{\sigma t}$ with $\sigma \in \mathbb{C}$, and assume $e_\sigma \in \mathcal{I}$,

Proposition 1.4: $Le_\sigma = H(\sigma)e_\sigma$ where $H(\sigma) = Le_\sigma(0)$

Proof:

$$\begin{aligned} L(\tau_\Delta e_\sigma) &= \tau_\Delta(Le_\sigma) \\ &= L(e^{-\sigma\Delta}e^{\sigma t}) = e^{-\sigma\Delta}Le_\sigma \end{aligned}$$

and thus

$$Le_\sigma = e^{\sigma\Delta}\tau_\Delta Le_\sigma$$

which implies

$$e^{-\sigma t}Le_\sigma = e^{-\sigma(t-\Delta)}\tau_\Delta$$

or equivalently

$$e_{-\sigma}Le_\sigma = \tau_\Delta(e_{-\sigma}Le_\sigma).$$

$e_{-\sigma}Le_\sigma$ is thus constant and hence $Le_\sigma = H(\sigma)e_\sigma$ where $H(\sigma) = Le_\sigma(0)$.

1.1.1 Formal Fourier analysis

Back to linearity: By definition, if

$$f = \sum_{k\in I} c_k e_{\sigma_k}$$

then

$$Lf = \sum_{k \in I} c_k H(\sigma_k) e_{\sigma_k}.$$

For all such functions, the response is entirely specified by the eigenvalues of e_σ with $\sigma \in \mathcal{D}$.

Extension: It is then tempting to extend this result to functions

$$f = \int_{\mathcal{D}} \widetilde{f}(\sigma) e_\sigma \mathrm{d}\sigma$$

for which one expects

$$Lf = \int_{\mathcal{D}} \widetilde{f}(\sigma) H(\sigma) e_\sigma \mathrm{d}\sigma.$$

Formal Fourier transform: The most classical case, Fourier transform, corresponds to $\mathcal{D} = i\mathbb{R}$. Indeed, formally, the inverse Fourier transform is given

$$f(t) = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{f}(\omega) e_{i\omega}(t) \mathrm{d}\omega$$

with the Fourier transform defined by

$$\widehat{f}(\omega) = \int_{\mathbb{R}} f(t) e_{-i\omega}(t) \mathrm{d}t$$

so that one expects

$$Lf = \frac{1}{2\pi} \int_{\mathbb{R}} H(i\omega) \widehat{f}(\omega) e_{i\omega}(t) \mathrm{d}\omega.$$

Laplace transform: $\mathcal{L}f : \sigma \mapsto \int_{\mathbb{R}} f(t) e_{-\sigma}(t) \mathrm{d}t$ with similar inversion formula

$$t \mapsto \frac{1}{2i\pi} \int_{\sigma \in a + i\mathbb{R}} \mathcal{L}f(\sigma) e^{\sigma t} d\sigma$$

could also be considered.

Transfer function: A LTI can be seen in the Fourier domain as the multiplication of the Fourier transform $\widehat{f}(\omega)$ by $\omega \mapsto H(i\omega)$, a function called the transfer function. We are *filtering* the frequencies by this function.

Again formally, if $Lf = h \star f$ then

$$\begin{aligned} Le_{i\omega}(t) &= \int_{\mathbb{R}} e_{i\omega}(u) h(t-u) \mathrm{d}u \\ &= \int_{\mathbb{R}} h(u) e_{-i\omega}(t-u) \mathrm{d}u = \widehat{h}(\omega) e_{i\omega} \end{aligned}$$

and thus

$$H(i\omega) = \widehat{h}(\omega).$$

We should work now to make those formal definitions mathematical ones... and as general as possible.

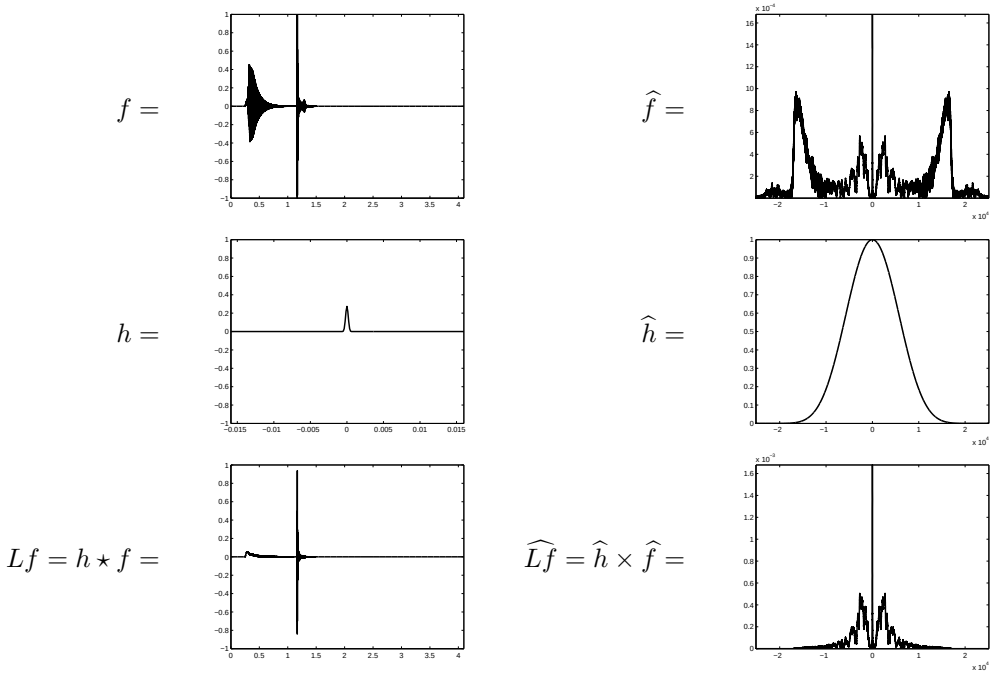


Figure 1.5: Frequential filtering

1.2 Fourier analysis

1.2.1 $\mathcal{S}$imple Fourier transform

The Schwartz class $\mathcal{S}$ is the class of $\mathcal{C}^\infty$ functions verifying

$$\sup_{\alpha\beta} \sup_{t\in\mathbb{R}} \left| t^\alpha \frac{\partial^\beta f}{\mathrm{d}t^\beta}(t) \right| < +\infty$$

Within this space, we say that $f_N \xrightarrow{\mathcal{S}} f$ if and only if

$$\lim_{N\to\infty} \sup_{\alpha\beta} \left| t^\alpha \frac{\partial^\beta (f_N - f)}{\mathrm{d}t^\beta}(t) \right| \to 0$$

Note, by definition, if $f \in \mathcal{S}$ then $\sup_{t\in\mathbb{R}} |(1+t^2)f(t)| = C < +\infty$ and thus $|f(t)| \leq C/(1+t^2)$ which implies that $f \in L^1$.

Fourier transform: As $\mathcal{F} \subset L^1$, the Fourier transform of $f \in \mathcal{S}$ can be defined by

$$\begin{aligned} \widehat{f}(\omega) &= \int_{\mathbb{R}} f(t) e_{-i\omega}(t) \mathrm{d}t \\ &= \int_{\mathbb{R}} f(t) e^{-i\omega t} \mathrm{d}t \end{aligned}$$

Beware: there are other definitions that differ either by a constant factor or a change of variable, $\omega = 2\pi f_r$ (angular frequency vs classical frequency).

Proposition 1.5 (Basic properties): If $f \in \mathcal{S}$

- $\widehat{\tau_\Delta f}(\omega) = e^{-i\omega\Delta}\widehat{f}(\omega)$
- $\widehat{e^{i\Delta t} f}(\omega) = \tau_\Delta \widehat{f}(\omega)$
- f real implies $\widehat{f}(-\omega) = \overline{\widehat{f}(\omega)}$

Proposition 1.6 (Regularity properties): If $f \in \mathcal{S}$

- $\widehat{\frac{\partial^p f}{\mathrm{d}t^p}}(\omega) = (i\omega)^p \widehat{f}(\omega)$
- $\widehat{-i^p t^p f}(\omega) = \frac{\partial^p \widehat{f}}{\mathrm{d}\omega^p}(\omega)$

Stability of $\mathcal{S}$: $f \in \mathcal{S} \implies \widehat{f} \in \mathcal{S}$

Convolution:

- If $h \in \mathcal{S}$ and $f \in \mathcal{S}$ then $h \star f \in \mathcal{S}$ and $\widehat{h \star f} = \widehat{h}\widehat{f}$.
- If $h \in \mathcal{S}$ and $f \in \mathcal{S}$ then $h \times f \in \mathcal{S}$ and $\widehat{h \times f} = \frac{1}{2\pi}\widehat{h} \star \widehat{f}$.

Inversion: If $f \in \mathcal{S}$ then

$$f(t) = \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(\omega) e^{i\omega} \mathrm{d}\omega$$

Remark: $2\pi f(t) = \widehat{\widehat{f}}(-t)$

Impulse response and transfer function: If $h \in \mathcal{S}$ and $f \in \mathcal{S}$ then

$$h \star f(t) = \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{h}(\omega)\widehat{f}(\omega) e^{i\omega} \mathrm{d}\omega$$

A *filter* is nothing but the multiplication by a transfer function in the Fourier domain...

Plancherel: If $f \in \mathcal{S}$ and $g \in \mathcal{S}$ then

$$\int_{\mathbb{R}} f(t)\overline{g(t)}\mathrm{d}t = \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(\omega)\overline{\widehat{g}(\omega)}\mathrm{d}\omega$$

1.2.2 3 *extensions*

Need for extensions: Previous analysis is valid only under the strong assumption that $f \in \mathcal{S}$, $h \in \mathcal{S}$. One would like to weaken this assumption... As a first step, we look at 3 *extensions* of the Fourier transform.

First *extension* is the extension to L^1, which is unfortunately not stable under the Fourier transform and leads to strong assumptions in order to obtain some results.

Second *extension* is the extension to L^2 using Plancherel equaity. L^2 is as $\mathcal{S}$ a stable class for the Fourier transform which plays a very important role, in particular in numerical analysis (Sobolev spaces).

Third extension is an extension to generalized functions $\mathcal{S}'$ (distributions) which is stable for the Fourier transform and the most appropriate setting to understand LTI operators.

Note that the section dedicatd to this extension is not intended as a comprehensive course but only as a (almost) mathematically correct introduction.

1.2.3 L^1 Fourier transform

L^1 Fourier transform: For $f \in L^1$,

$$\widehat{f}(\omega) = \int_{\mathbb{R}} f(t) e_{-i\omega}(t) \mathrm{d}t$$

L^1 continuity (in $\mathcal{S}$): For any $f \in L^1$ (or $\mathcal{S}$)

$$\|\widehat{f}\|_\infty \leq |f|_1$$

The density of $\mathcal{S}$ in L^1 allows to see the Fourier transform in L^1 as the extension by continuity of the Fourier transform in $\mathcal{S}$...

Proposition 1.7: If $f \in L^1$ then $\widehat{f}$ is well defined and is a bounded continuous function which vanishes at infinity.

Proposition 1.8 (Basic properties): If $f \in L^1$,

- $\widehat{\tau_\Delta f}(\omega) = e^{-i\omega\Delta}\widehat{f}(\omega)$
- $\widehat{e^{i\Delta t} f}(\omega) = \tau_\Delta \widehat{f}(\omega)$
- f real implies $\widehat{f}(-\omega) = \overline{\widehat{f}(\omega)}$

Proposition 1.9 (Regularity properties):

- If $\frac{\partial^p f}{\mathrm{d}t^p} \in L^1$ then $\widehat{\frac{\partial^p f}{\mathrm{d}t^p}}(\omega) = (i\omega)^p \widehat{f}(\omega)$
- If $t^p f \in L^1$ then $\widehat{-i^p t^p f}(\omega) = \frac{\partial^p \widehat{f}}{\mathrm{d}\omega^p}(\omega)$

Regularity is *almost* equivalent to fast decay of the Fourier transform.

Proposition 1.10: If $f \in L^1$ and $\widehat{f} \in L^1$ then

$$f(t) = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{f}(\omega) e^{i\omega t} \mathrm{d}\omega.$$

This property could be rewritten as $\widehat{\widehat{f}} = 2\pi f(-.)$

Proposition 1.11: If $f \in L^1$ and $h \in L^1$ then $Lf = h \star f \in L^1$ and

$$\widehat{Lf}(\omega) = \widehat{h}(\omega) \times \widehat{f}(\omega).$$

Proposition 1.12: If $\widehat{f} \in L^1, \widehat{h} \in L^1$ and $f \times h \in L^1$ then

$$\widehat{f \times h}(\omega) = \frac{1}{2\pi} \widehat{h} \star \widehat{f}(\omega)$$

Impulse response and transfer function: If $f \in L^1$, $h \in L^1$ and $\widehat{h} \times \widehat{f} \in L^1$, we can thus write

$$Lf(t) = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{h}(\omega) \widehat{f}(\omega) e_{i\omega}(t) d\omega$$

1.2.4 L^2 extension

L^2 continuity in $\mathcal{S}$: By Plancherel equality, for all $f \in \mathcal{S}$,

$$\|f\|_2^2 = \frac{1}{2\pi} \|\widehat{f}\|_2^2.$$

The density of $\mathcal{S}$ in L^2 allows to extend $f \mapsto \hat{f}$ to L^2. Note that there is no explicit formula for the Fourier transform. To stress this, we will use the notation $\mathcal{F}f$ instead of $\widehat{f}$ when $f \notin L^1$.

Stability of L^2: If $f \in L^2$ then $\mathcal{F}f \in L^2$.

Limit formula: If $f \in L^2$,

$$\mathcal{F}f \overset{L^2}{=} \lim_{M \to \infty} \omega \mapsto \int_{-M}^{M} f(t) e^{-i\omega t} \mathrm{d}t.$$

Proposition 1.13 (Basic properties): If $f \in L^2$,

- $\mathcal{F}(\tau_\Delta f) = e^{-i\omega\Delta} \mathcal{F}f$
- $\mathcal{F}(e^{i\Delta t} f) = \tau_\Delta \mathcal{F}f$
- f real implies $\mathcal{F}f(-\cdot) = \overline{\mathcal{F}f}$

Proposition 1.14 (Regularity properties):

- If $\frac{\partial^p f}{\mathrm{d}t^p} \in L^2$ then $\mathcal{F}\left(\frac{\partial^p f}{\mathrm{d}t^p}\right) = (i\omega)^p \mathcal{F}f$
- If $t^p f \in L^2$ then $\mathcal{F}(-i^p t^p f) = \frac{\partial^p \mathcal{F}f}{\mathrm{d}\omega^p}$

Again, regularity is *almost* equivalent to fast decay of the Fourier transform.

Proposition 1.15: If $f \in L^2$, $f = \frac{1}{2\pi}\mathcal{F}\mathcal{F}(-\cdot)$ or equivalently

$$f \overset{L^2}{=} \lim_{N\to\infty} \omega \mapsto \frac{1}{2\pi}\int_{-N}^{N} \mathcal{F}f(\omega)e^{i\omega t}\mathrm{d}\omega.$$

We denote by $\mathcal{F}^{-1}$ this application, by construction $\mathcal{F}^{-1}(U) = \frac{1}{2\pi}\mathcal{F}U(-\cdot)$.

Proposition 1.16: If $f \in L^2$ and $h \in L^1$ then $Lf = h \star f \in L^2$ and

$$\mathcal{F}(Lf) = \widehat{h} \times \mathcal{F}f.$$

Proposition 1.17: If $f \in L^2, h \in L^2$ and $\mathcal{F}f \times \mathcal{F}h \in L^2$ then $Lf = h \star f \in L^2$ and

$$\mathcal{F}(Lf) = \mathcal{F}h \times \mathcal{F}f$$

Using the stability of L^2, we obtain

Proposition 1.18: If $f \in L^2, h \in L^2$ and $\mathcal{F}h \star \mathcal{F}f \in L^2$ then

$$\mathcal{F}(h \times f) = \mathcal{F}h \star \mathcal{F}f$$

Impulse response and transfer function: If $f \in L^2$, $h \in L^1$, we can thus write $Lf = \mathcal{F}^{-1}(\widehat{h} \times \mathcal{F}f)$

If $f \in L^2$, $h \in L^2$ and $\mathcal{F}h \times \mathcal{F}f \in L^2$, we obtain $Lf = \mathcal{F}^{-1}(\mathcal{F}h \times \mathcal{F}f)$

1.2.5 Dirac delta function and distributions

Dirac delta function: Originally the distributions did not appear to extend the Fourier transform but rather to give a proper framework to work with Dirac delta *functions*.

The Dirac delta function at 0 is the operator that associates to a continuous function f its value at 0, $f(0)$. Although this operator cannot be written as a *scalar* product with a function δ

$$f \mapsto \int_{\mathbb{R}} f(t)\delta(t)\mathrm{d}t,$$

it can be written as a limit of such scalar product

$$f \mapsto \int_{\mathbb{R}} f(t)K_n(t)\mathrm{d}t$$

where K_n is a sequence of approximate identity. In the distribution theory, one identify the function K_n with the corresponding operator on continuous function. Its action on a function f is denoted $\langle K_n, f\rangle$ to stress its resemblance with a scalar product. If we denote by $\langle \delta, f\rangle$ the action of δ on f, i.e. $f(0)$, we observe that

$$\forall f \mathcal{C}^0, \langle K_n, f\rangle \to \langle \delta, f\rangle.$$

In the distribution theory such a property will correspond to the fact that $K_n \to \delta$...

Distributions: More precisely, a set of Distributions is defined as the set of Continuous Linear Forms on a (very) regular function set. The classical distributions $\mathcal{D}'$ are defined by their actions on $\mathcal{D}$ the set of compactly supported $\mathcal{C}^\infty$ functions while the tempered distribution $\mathcal{S}'$ are defined by their actions on $\mathcal{S}$. The continuity assumption means that if $\phi_n \to \phi$ in $\mathcal{D}$ (respectively in $\mathcal{S}$) and U belongs to $\mathcal{D}'$ (respectively to $\mathcal{S}'$) then $\langle U, \phi_n\rangle \to \langle U, \phi\rangle$. The $'$ in the notation stresses that those spaces are topological dual (for which the topology is implicitely specified through the previous sequential definition...).

One verify that $\mathcal{S}' \subset \mathcal{D}'$. One can verify that $L^p \subset \mathcal{S}'$. Finally $\mathcal{C}^\infty \not\subset \mathcal{S}'$ but $\mathcal{C}^\infty \subset \mathcal{D}'$.

By construction, $\delta = \delta_0$ belongs to $\mathcal{S}' \subset \mathcal{D}'$.

Properties: Without any proofs, we give some important properties of distributions

- Translation invariance of $\mathcal{S}'$ and $\mathcal{D}'$.
- Differentiation for $U \in \mathcal{D}'$: U' always exists and is defined by $\langle U', \phi\rangle = -\langle U, \phi'\rangle$ (By parts integration)
- Fourier transform for $U \in \mathcal{S}'$: the Fourier transform $\mathcal{F}U$ always exists and is defined by $\langle \mathcal{F}U, \phi\rangle = \langle U, \widehat{\phi}\rangle$ (Plancherel)
- Inverse Fourier transform for $U \in \mathcal{S}'$: $\mathcal{F}^{-1}U = \frac{1}{2\pi}\mathcal{F}U(-\cdot)$.
- Convolution: $\mathcal{F}(U \star V) = \mathcal{F}U \times \mathcal{F}V$ (if everything is well defined which is not always the case...)

Fourier series: Let f be a T-periodic $\mathcal{C}^1$ function, f can be decomposed into its Fourier series

$$f = \sum_{n\in\mathbb{Z}} c_n e^{in\frac{2\pi}{T}}$$

where the Fourier coefficients c_n are defined by

$$c_n = \frac{1}{T}\int_{-T/2}^{T/2} f(t)e^{-in\frac{2\pi}{T}t}\mathrm{d}t.$$

With our strong regularity assumption, the convergence is true pointiwse or in L^2-loc for instance.

Approximation of the identity: Let $\chi_{[0,T]}$ be a compactly supported non negative function such that

$$\sum_{k\in\mathbb{Z}} \chi_{[0,T]}(\cdot - kT) = 1,$$

$$c_n = \frac{1}{T}\int_{\mathbb{R}} f(t)\chi_{[0,T]}(t)e^{-in\frac{2\pi}{T}t}\mathrm{d}t.$$

Such functions $\chi_{[0,T]}$ exist and can even be chosen in $\mathcal{C}^\infty$.

Fourier series for periodic distribution: Let U be a T-periodic distribution in $\mathcal{D}'$.

$$U = \sum_{n\in\mathbb{Z}} c_n e_{in\frac{2\pi}{T}}$$

with

$$c_n = \langle U, \frac{\chi_{[0,T]}}{T} e_{-in\frac{2\pi}{T}}\rangle.$$

Furthermore U belongs to $\mathcal{S}'$ and c_n grows at most polynomialy with n. Its Fourier transform exists and is given by

$$\mathcal{F}U = 2\pi\sum_n c_n \delta_{n\frac{2\pi}{T}}.$$

Examples:

- Box function and sinc function:

$$\begin{aligned}
\widehat{\mathbf{1}_{[-\Delta,\Delta]}}(\omega) &= \int_{-\Delta}^{\Delta} e^{-i\omega t}\mathrm{d}t = \left[\frac{e^{-i\omega t}}{-i\omega}\right]_{-\Delta}^{\Delta} \\
&= \frac{-e^{-i\omega\Delta} + e^{i\omega\Delta}}{i\omega} = \Delta\frac{\sin\Delta\omega}{\Delta\omega} \\
&= \Delta\,\mathrm{sinc}(\Delta\omega) \\
\mathcal{F}\,\mathrm{sinc}(\Delta\cdot) &= \frac{1}{\Delta}\mathcal{F}\mathcal{F}\mathbf{1}_{[-\Delta,\Delta]} \\
&= \frac{2\pi}{\Delta}\mathbf{1}_{[-\Delta,\Delta]}
\end{aligned}$$

- Dirac and complex exponential:

$$\begin{aligned}
\langle \mathcal{F}\delta_\tau, \phi\rangle = \langle \delta_\tau, \widehat{\phi}\rangle = \widehat{\phi}(\tau) &= \int_{\mathbb{R}} \phi(t)e^{-i\tau t}\mathrm{d}t \\
\mathcal{F}\delta_\tau &= e_{-i\tau} \\
\mathcal{F}e_{i\omega_0} &= \mathcal{F}\mathcal{F}\delta_{-\omega_0} \\
&= 2\pi\delta_{\omega_0}
\end{aligned}$$

- Dirac comb: $\sum_{n\in\mathbb{Z}} \delta_{n\Delta}$

$$\begin{aligned}\mathcal{F}\left(\sum_{n\in\mathbb{Z}} \delta_{n\Delta}\right) &= \sum_{n\in\mathbb{Z}} \mathcal{F}\delta_{n\Delta} \\ &= \sum_{n\in\mathbb{Z}} e^{in\delta}\end{aligned}$$

Proposition 1.19 (Poisson formula): $\mathcal{F}\left(\sum_{n\in\mathbb{Z}} \delta_{n\Delta}\right) = \frac{2\pi}{\Delta}\sum_{n\in\mathbb{Z}} \delta_{n2\pi/\Delta}$

Proof: As $C = \sum_{n\in\mathbb{Z}} \delta_{n\Delta}$ is a Δ periodic distribution

$$C = \sum_{n\in\mathbb{Z}} c_n e_{in2\pi/\Delta}$$

with $c_n = \langle C, \frac{1}{\Delta}\chi_{[0,\Delta]} e_{-in2\pi/\Delta}\rangle = \frac{1}{\Delta}$ and thus

$$C = \frac{1}{\Delta}\sum e_{in2\pi/\Delta}.$$

As $\mathcal{F}e_{in2\pi/\Delta} = 2\pi\delta_{n2\pi/\Delta}$,

$$\mathcal{F}C = \frac{2\pi}{\Delta}\sum_{n\in\mathbb{Z}} \delta_{n2\pi/\Delta}$$

LTI and distributions: Most general definition of LTI operator as operator acting on distributions...

Impulse response and transfer function: If $LU = h \star U$ (and everything is well defined) then

$$h = h \star \delta = L\delta$$

hence the name impulse response and

$$\langle LU, \phi\rangle = \langle LU, \widehat{\mathcal{F}^{-1}\phi}\rangle = \langle \mathcal{F}(LU), \mathcal{F}^{-1}\phi\rangle = \langle \mathcal{F}^{-1}(\mathcal{F}h \times \mathcal{F}U), \phi\rangle$$

where we recover the spectral representation of the filtering.

One can verify that this interpretation is thus coherent with the one of the previous section.

1.3 LTI, convolution systems and examples

1.3.1 LTI and convolution system

a LTI operator is not necessarily a convolution: A folk's theorem states that every LTI is a convolution system (this can be found in many courses). This not true!

BIBO counter-example: $U \mapsto (t \mapsto \lim_{\Delta\to\infty}\langle U, \phi_\Delta(\cdot - t)\rangle)$ where $\phi_\Delta = \frac{1}{\Delta}\phi(t/\Delta)$ with $\phi \in \mathcal{D}$, $\phi \geq 0$ verify $L\delta = 0$ while $L\mathbf{1} = \int \phi(f)\mathrm{d}t\mathbf{1}$.

Issue lack of continuity in $\mathcal{D}'$...

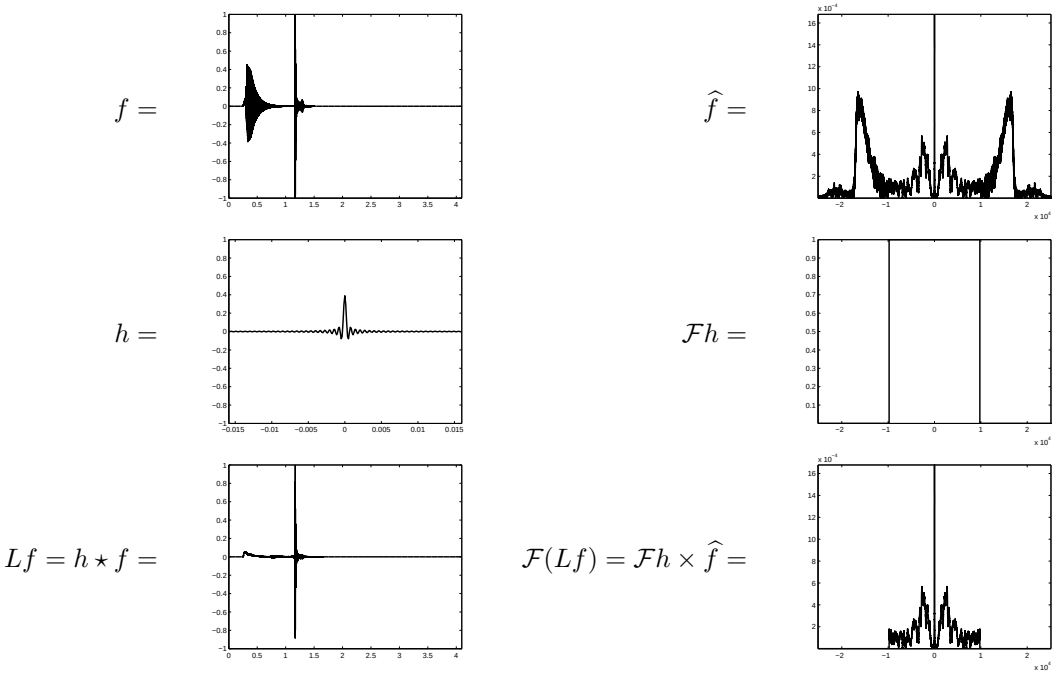


Figure 1.6: Frequential filtering II

LTI and convolution: Most general results:

- If L is continuous from $\mathcal{D}$ into $\mathcal{D}'$ then it exists an impulse response $h \in \mathcal{D}'$ such that for $f \in \mathcal{D}, Lf = h \star f$
- If L is furthermore continuous from $\mathcal{D}'$ (or $\mathcal{S}'$) into $\mathcal{D}'$ then $L\delta = h$ and L is entirely specified by its impulse response h.

In practice, almost every LTI can be rewritten as a convolution...

1.3.2 Application to Amplitude Modulation multiplexing

AM multiplexing: To illustrate the power of those tools, we describe how to to use them to analyze a classical example, the AM multiplexing system used in radio transmission. Assume we have at hand K signals s_k (radio shows) each having a Fourier transform supported in $[-B, B]$, we want to mix them into a single signal S to be transmitted in such a way that all those sigals can be recovered from S.

Multiplexing: We start by the following observation, if $\mathcal{F}s_k$ is supported in $[-B, B]$ then $\mathcal{F}(s_k \times \cos(\omega_k t)) = \mathcal{F}s_k \star 1/2(\delta_{-\omega_k} + \delta_{\omega_k})$ is supported in $[-\omega_k - B, -\omega_k + B] \cup [\omega_k - B, \omega_k + B]$.

Thus if we let $\omega_k = 2(k-1)B$, all the signals $s_k \times \cos(\omega_k t)$ have disjoint frequency support. We can hope thus to recover them from their sum $S = \sum_{k=1}^{K} s_k \times \cos(\omega_k t)$.

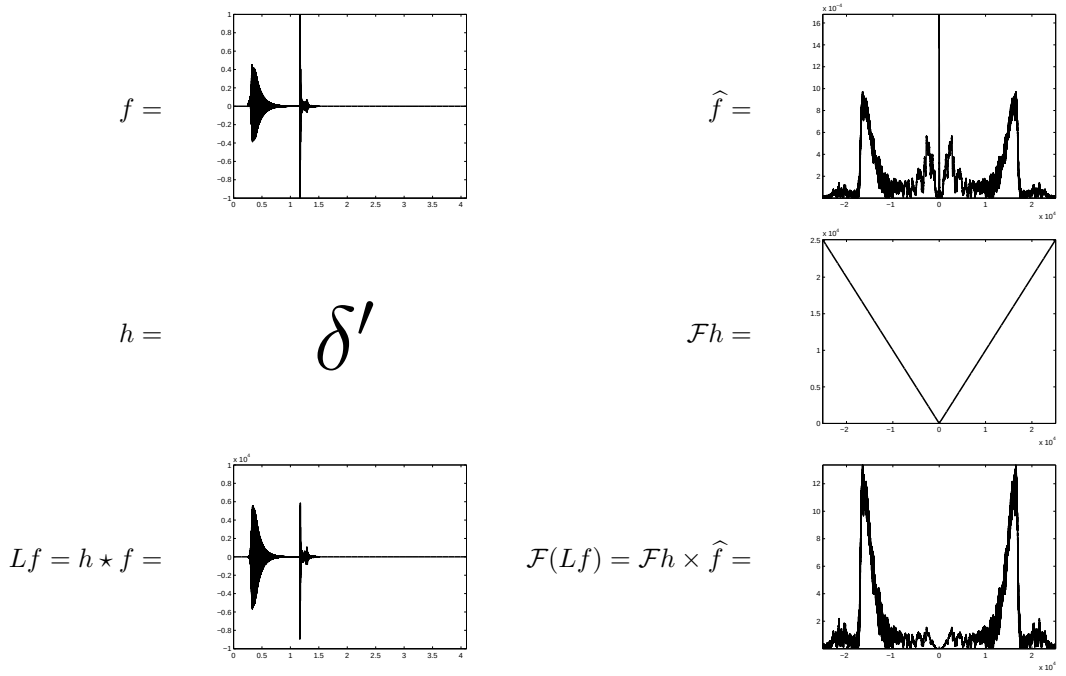


Figure 1.7: Frequential filtering III

Demultiplexing: Indeed, if one multiplies S by $\cos(\omega_{k'}t)$, one obtains

$$S \times \cos(\omega_{k'}t) = \sum_{k=1}^{K} s_k \times \cos(\omega_k t) \times \cos(\omega_{k'}t)$$

Now

$$\begin{aligned}\mathcal{F}(s_k \times \cos(\omega_k t) \times \cos(\omega_{k'}t)) &= \mathcal{F}s_k \star 1/2(\delta_{-\omega_k} + \delta_{\omega_k}) \star 1/2(\delta_{-\omega_{k'}} + \delta_{\omega_{k'}}) \\ &= \frac{1}{4}\mathcal{F}s_k \star (\delta_{-\omega_k-\omega'_k} + \delta_{-\omega_k+\omega'_k} + \delta_{\omega_k-\omega'_k} + \delta_{\omega_k+\omega'_k})\end{aligned}$$

is such that its support does not intersect $[-B, B]$ as long as $k \neq k'$ and in this case

$$\mathcal{F}(s_k \times \cos(\omega_k t) \times \cos(\omega_k t)) = \frac{1}{4}\mathcal{F}(s_k) \star (2\delta + \delta_{-2\omega_k} + \delta_{2\omega_k}).$$

If we let h be the low pass filter such that $\mathcal{F}h_B = \mathbf{1}_{[-B,B]}$, we obtain thus that

$$\mathcal{F}(S \times \cos(\omega_k t)) \times \mathcal{F}h = \frac{1}{2}\mathcal{F}s_k$$

or equivalently in the spatial domain

$$(S \times \cos(\omega_k t)) \star h_B = \frac{1}{2}s_k.$$

Remark that our process is not LTI as the multiplication by $\cos(\omega_k t)$ is not a spatial LTI (but it is a frequential one).

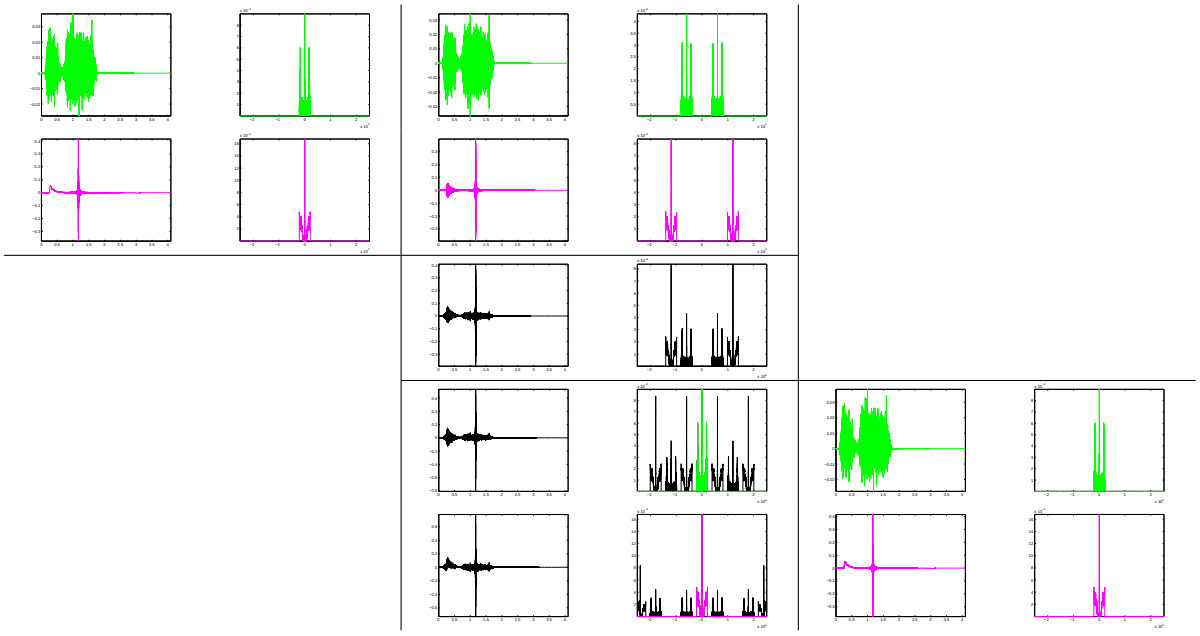


Figure 1.8: AM Multiplexing/Demultiplexing

So far, we did not discuss the possibility to implement the filter used. In practice, this issue is however very important as those analog system are implement using real hardware...

1.3.3 Causality and realizable filter

Causality: The first observation is that, as one cannot foresee the future, any physical system should be a causal one. This turns out to be an issue as, for instance, the ideal pass-band filter is not causal...

Electrical implementation: Classical implementation rely on electrical circuit made of Resistor (R), Inductor (L), Capacitor (C) and Operator Amplifier (OA) or rather in integrated circuit (SSI, MSI, LSI, VLSI, ULSI) only with R, C and OA for sake of space. As you may recall, in an electrical circuit, the input f and the output Lf are relied through a linear differential equation

$$\sum_{k=0}^{K_a} a_k f^{(k)} = \sum_{k=0}^{K_b} b_k (Lf)^{(k)}$$

which can be conveniently rewritten in the Fourier domain as

$$\sum_{k=0}^{K_a} a_k (i\omega)^k \mathcal{F}f = \sum_{k=0}^{K_b} b_k (i\omega)^k \mathcal{F}(Lf)$$

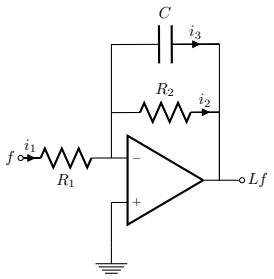
so that

$$\mathcal{F}(Lf) = \frac{\sum_{k=0}^{K_a} a_k (i\omega)^k}{\sum_{k=0}^{K_b} b_k (i\omega)^k} \mathcal{F}f$$

and the transfer function is rational in $i\omega$.

BIBO stability implies that $K_b \geq K_a$ and that the poles of the rational fraction are in the right hand side of the plane.

Proposition 1.20: If h is a rational function then there exist a electrical circuit implementing a filter g such that $|\mathcal{F}g| = |\mathcal{F}h|$ as soon as $K_b \geq K_a$.



Example of a low pass filter: An electrical circuit analysis yields:

$$f = R_1 i_1, \quad Lf = -R_2 i_1, \quad Lf' = -C i_3, \text{ and } \quad i_1 = i_2 + i_3.$$

In the Fourier domain, we have thus

$$\widehat{f} = R_1 \widehat{i_1}, \quad \widehat{Lf} = -R_2 \widehat{i_2}, \quad \omega \widehat{Lf} = -C \widehat{i_3}, \text{ and } \quad \widehat{i_1} = \widehat{i_2} + \widehat{i_3}.$$

A straightforward computation allows to obtain the transfer function $H(i\omega)$ such that $\widehat{Lf} = H(i\omega)\widehat{f}$:

$$H(i\omega) = \frac{-R_2/R_1}{1 + \frac{R_2}{C} i\omega}$$

If $R_1 = R_2$,

$$|H(i\omega)| = \frac{1}{\sqrt{1 + \left(\frac{R_2}{C}\omega\right)^2}}$$

this is a low pass filtering with cutoff frequency $\omega_C \sim C/R_2$!

Better (rational) approximation of the box function could be obtained with more complex circuits... (this is the subject of analog filter design theory)

Chapter 2

Discrete Signal Processing

In the previous section, we have studied analog signal which are parts of our environment. However, we are living in a more and more digital world and in this world signals are not continuous but discrete, they are sequences of numbers. To be more accurate, as a computer can only handle finite precision, those numbers are quantized so they belong to a finite set. In this section, we should neglect this quantization.

2.1 Discretization and sampling

2.1.1 Discretization and Fourier transform

AD conversion: We consider here a canonical process of Analog to Digital and Digital to Analog conversions from a theoretical point of view, that will lead to practical insight.

Sampling: We assume the AD conversion is the most simple one. Let $\Delta \in \mathbb{R}_+$ be a discretization step, we assume we can acquire from a function f the sequence of samples

$$(f(n\Delta))_{n\in\mathbb{Z}}\,.$$

Band-limited functions: Note that we should assume that f is at least continuous in order to define the samples. In order to derive our theory, we will require a much stronger regularity

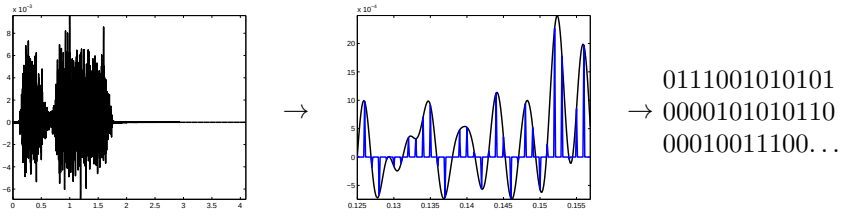


Figure 2.1: Digital world

assumption. We will assume that the Fourier transform $\mathcal{F}f$ of f exists (which implies that f is at most polynomially increasing) and that has a compact support. We will call such a function a band-limited function. This is a very strong assumption as one can show that this implies that f is analytic.

The sequence $(f(n\Delta))_{n\in\mathbb{Z}}$ is not the most interesting way to represent the discretized version of f. Within the distribution setting, one can define a better representation

$$f_\Delta = \Delta \sum_n f(n\Delta)\delta_{n\Delta}$$

which converge to f in $\mathcal{D}'$ when Δ goes to 0.

As soon as $f(n\Delta)$ is at most polynomially increasing, f_Δ belongs to $\mathcal{S}'$ and one can thus compute its Fourier transform. It turns out that if f is band-limited then $\mathcal{F}f_\Delta$ is the periodization of $\mathcal{F}f$ with period $2\pi/\Delta$.

Theorem 2.1: if f is band-limited then

$$\mathcal{F}f_\Delta = \sum_n \mathcal{F}f(\cdot - n2\pi/\Delta)$$

Proof: A proof of this result can be obtained easily by noticing that

$$f_\Delta = \Delta f \times \sum_{n\in\mathbb{Z}} \delta_{n\Delta}$$

so that

$$\begin{aligned}\mathcal{F}f_\Delta &= \frac{1}{2\pi}\Delta\mathcal{F}(\sum_{n\in\mathbb{Z}}\delta_{n\Delta}) \star \mathcal{F}f \\ &= \sum_{n\in\mathbb{Z}} \delta_{n2\pi/\Delta} \star \mathcal{F}f \\ &= \sum_n \mathcal{F}f(\cdot - n2\pi/\Delta)\end{aligned}$$

The band-limited assumption implies that $\mathcal{F}f$ is a compactly supported distribution and then that all the convolution make sense.

2.1.2 The Shannon sampling theorem

Formal derivation: If $\mathcal{F}f$ is supported in $(-\pi/\Delta, \pi/\Delta)$ then there is no overlapping in the periodization process. Formally, we deduce

$$\mathcal{F}f = \mathcal{F}f_\Delta \times \mathbf{1}_{(-\pi/\Delta,\pi/\Delta]}$$

and then

$$\begin{aligned}f &= f_\Delta \star \frac{1}{\Delta}\operatorname{sinc}(\cdot/\Delta) \\ &= \sum_{n\in\mathbb{Z}} f(n\Delta)\operatorname{sinc}\left(\frac{\cdot - n\Delta}{\Delta}\right)\end{aligned}$$

which gives a reconstruction formula of f from its samples. There is however a subtle issue in this reasoning: the object $\mathcal{F}f_\Delta \times \mathbf{1}_{(-\pi/\Delta,\pi/\Delta]}$ is not defined in term of distribution due to the lack of regularity of the characteristic function.

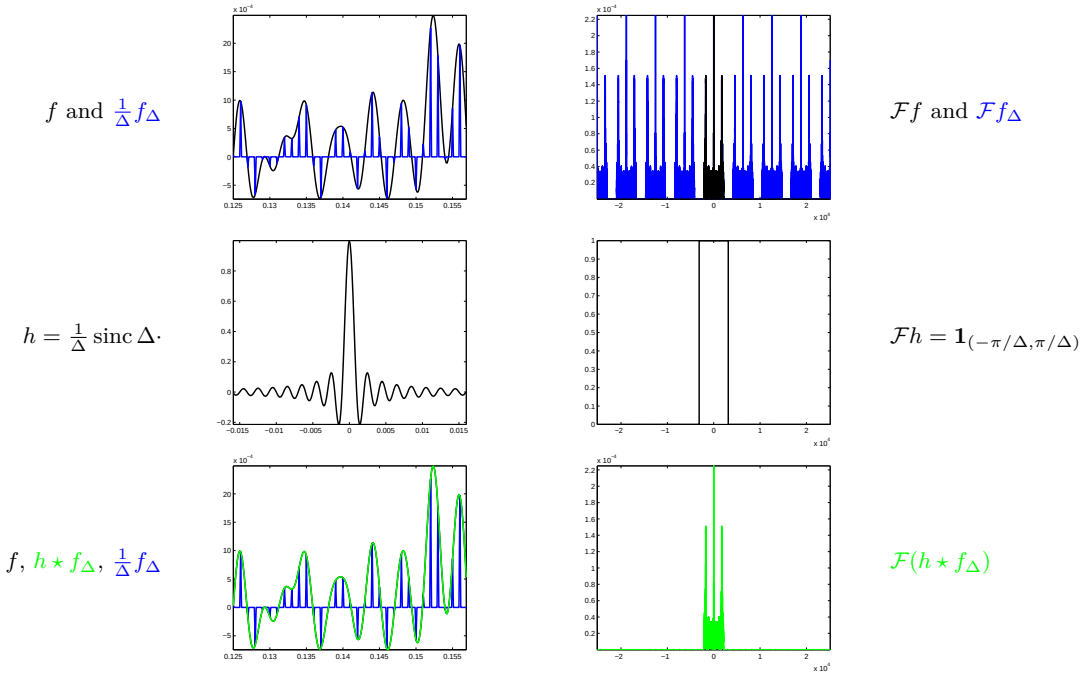


Figure 2.2: Shannon Theorem

A first fix is to reinforce the support assumption into supported in $[(-(1-\epsilon)\pi/\Delta, (1-\epsilon)\pi/\Delta]$ and replacing the characteristic function by a regularized version h_Δ with value 1 in $[-(1-\epsilon)\pi/\Delta, (1-\epsilon)\pi/\Delta]$. This will indeed lead to a reconstruction formula using the same derivation.

A second one: If we want to keep the assumption "supported in $(-\pi/\Delta, \pi/\Delta)$", then we need an additional assumption that could be either $f \in L^2$ or f is a finite sum of complex exponential $\sum_{k=1} K a_k e^{i\omega_k t}$.

Theorem 2.2 (Shannon, Nyquist, ...): If f is a band-limited function supported in $(-\pi/\Delta, \pi/\Delta)$ and either $f \in L^2$ or f is a finite sum of complex exponential (or a sum of two such functions) then

$$f(t) = \sum_{n\in\mathbb{Z}} f(n\Delta)\operatorname{sinc}\left(\frac{t-n\Delta}{\Delta}\right)$$

The two proofs are very different...

2.1.3 Aliasing

A natural question is what is going on if one sample at rate Δ a function f which is not band-limited in $(-\pi/\Delta, \pi/\Delta)$.

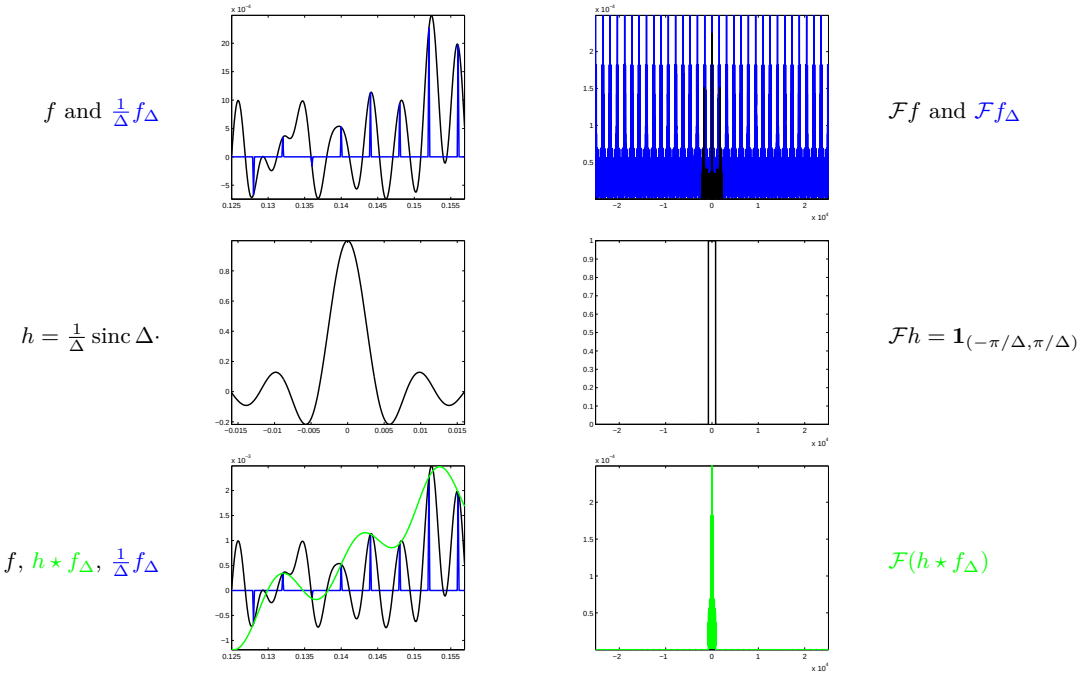


Figure 2.3: Aliasing

Aliasing: If f is band-limited, one still has

$$\mathcal{F}f_\Delta = \sum_n \mathcal{F}f(\cdot - n2\pi/\Delta)$$

but there is an overlap between the different translation of the support $[-B, B]$ of $\mathcal{F}f$ so that $\mathcal{F}f$ cannot be recovered from $\mathcal{F}f_\Delta \times \mathbf{1}_{[-\pi/\Delta,\pi/\Delta]}$. Furthermore, the Fourier transform of the reconstruction is different from $\mathcal{F}f \times \mathbf{1}_{[-\pi/\Delta,\pi/\Delta]}$. This phenomena is called aliasing.

The cos **case:** An interesting insight is obtained by studying the case $f = \cos(\omega_0 t) = \frac{1}{2}(e^{-i\omega_0 t} + e^{i\omega_0 t})$ which is a finite sum of complex exponentials. One has thus

$$\mathcal{F}f = \pi(\delta_{-\omega_0} + \delta_{\omega_0})$$

and, as f is band-limited,

$$\mathcal{F}f_\Delta = \pi \sum_{n\in\mathbb{Z}} (\delta_{-\omega_0 + n2\pi/\Delta} + \delta_{\omega_0 + n2\pi/\Delta}).$$

One verify then that

$$\mathcal{F}f_\Delta \times \mathbf{1}_{[-\pi/\Delta,\pi/\Delta]} = \pi(\delta_{-\omega_0'} + \delta_{\omega_0'})$$

where ω_0' is the unique translate $\omega_0 + n2\pi/\Delta$ in $[-\pi/\Delta, \pi/\Delta)$ (i.e. $\omega_0' = (\omega_0 + \pi/\Delta) \mod 2\pi/\Delta - \pi/\Delta$). As soon as, $\omega_0 \notin [-\pi/\Delta, \pi/\Delta), \omega_0 \neq \omega_0'$ and thus we do not recover f. For instance if $\omega_0 = \pi/\Delta + \theta$ with $0 < \theta < \pi/\Delta$, $\omega_0' = -\pi/\Delta + \theta$ so that the angular speed of the reconstruction is $\pi/\Delta - \theta$ which is slower than the original speed.

This phenomenon can be observed in practice in movies in which wheels sometime appear to rotate in the reverse way that they should. This is due to the time sampling of the movie.

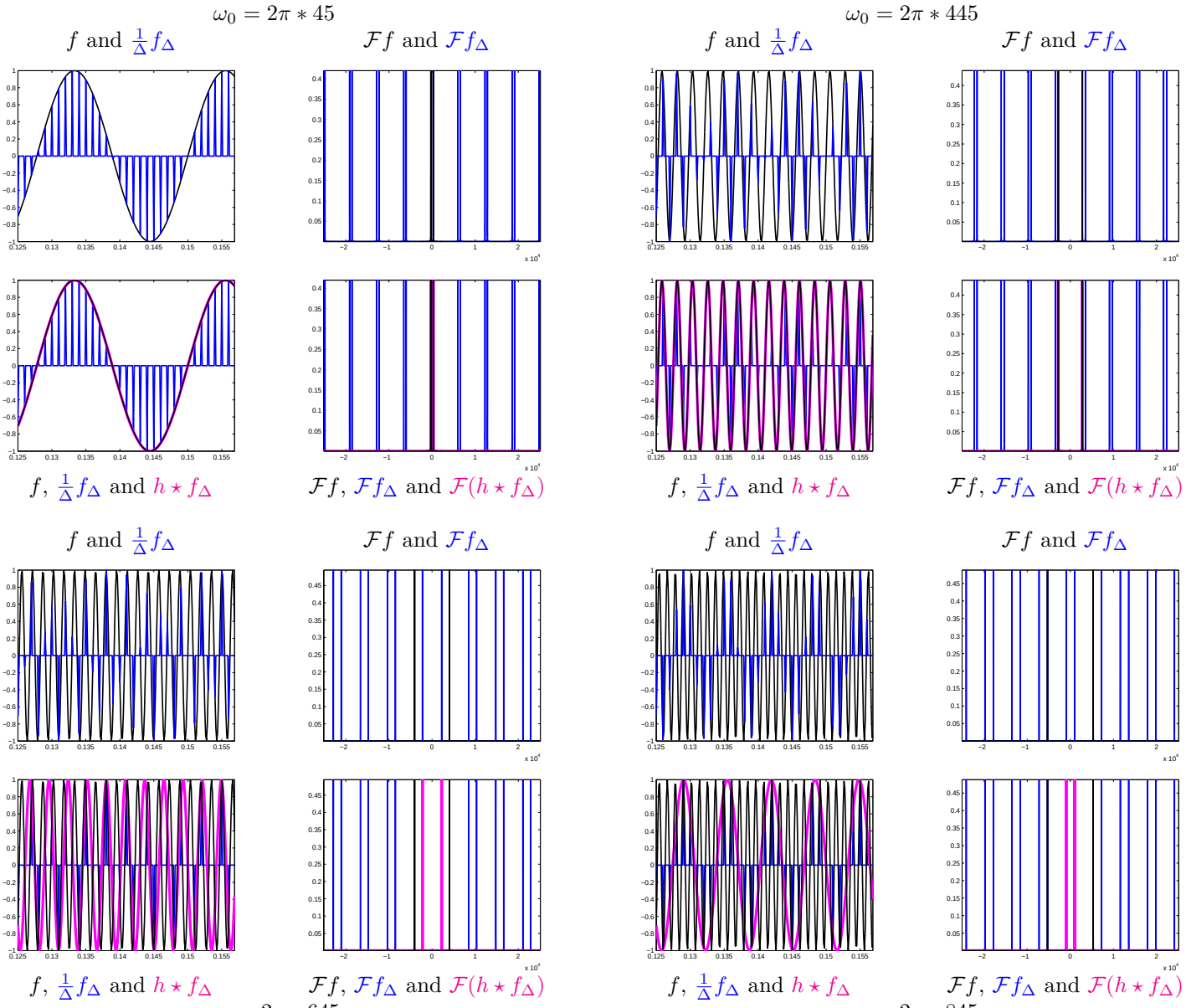


Figure 2.4: The cos case

Shannon formula stability: If f is not band-limited but satisfies some structural assumption, one can prove the stability of this sampling process, which is only mildly perturbed if there is only a small component outside the set $[-\pi/\Delta, \pi/\Delta]$

Theorem 2.3 (Brown): If $f \in L^2$ is continuous and such that $\mathcal{F}f \in L^1$ then

$$\left| f(t) - \sum_{n\in\mathbb{Z}} f(n\Delta)\,\mathrm{sinc}\left(\frac{t-n\Delta}{\Delta}\right)\right| \leq 2\int_{|\omega|>\pi/\Delta} |\mathcal{F}f(\omega)|\mathrm{d}\omega$$

Note that essentially the same results holds for finite sum of exponential if we replace the right hand side by

$$2\sum_{k,|\omega_k|\geq\pi/\Delta} |a_k|$$

Optimized AD conversion: Note that whatever g

$$\mathcal{F}\left(\sum_{n\in\mathbb{Z}} g(n\Delta)\,\mathrm{sinc}\left(\frac{t-n\Delta}{\Delta}\right)\right)$$

is compactly supported in $[-\pi/\Delta, \pi/\Delta]$ if it exists and if $f \in L^2$ then

$$\arg\min_{g\in L^2, \mathrm{supp}\mathcal{F}g\subset[-\pi/\Delta,\pi/\delta]} \|f-g\|^2 = \mathcal{F}^{-1}\left(\mathcal{F}f \times \mathbf{1}_{[-\pi/\Delta,\pi/\Delta]}\right).$$

The best L^2 AD strategy is thus to project f into the space of band-limited function supported in $[-\pi/\Delta, \pi/\Delta]$ by low pass filtering and to sample the resulting function.

Smoothing before sampling is the strategy always used in practice.

2.2 Discrete Signal and LTI

2.2.1 Discrete Signal and LTI

Discrete signals are nothing but sequences $(\mathrm{f}[n])_{n\in\mathbb{Z}}$.

LTI operators are (continuous) Linear Translation Invariant operators on sequences.

Kronecker symbol: δ is the sequence $\delta[0] = 1$ and $\delta[n] = 0$ for $n \neq 0$ and, as for the Dirac delta functions, let $\delta_k = \delta[\cdot - k]$.

Kronecker decomposition: $\mathrm{f}[n] = \sum_k \mathrm{f}[k]\delta[n-k] = \sum_k \mathrm{f}[h]\delta_k[n]$

Impulse response: $\mathrm{h} = \mathrm{L}\delta$

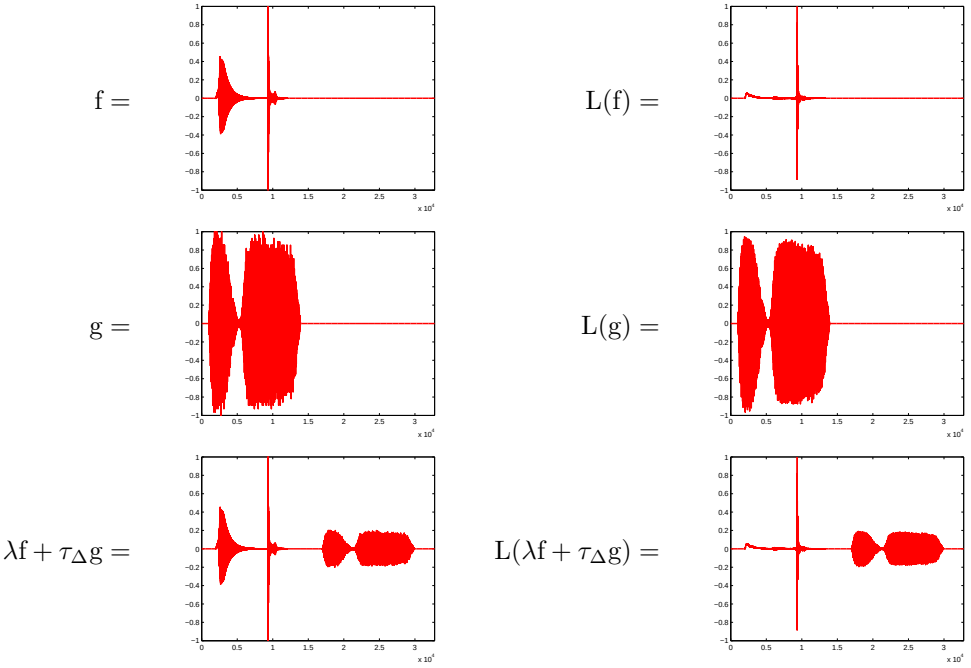


Figure 2.5: LTI

Convolution: Formally:

$$\mathrm{Lf} = \mathrm{L}(\sum_k \mathrm{f}[k]\delta_k) = \sum_k \mathrm{f}[k]\mathrm{L}\delta[\cdot - k] = \mathrm{h} \star \mathrm{f}$$

As in the analog case, this holds under some mild continuity assumptions on L (and always holds for finitely supported signals).

Stability: is again nothing but continuity for specific norm.

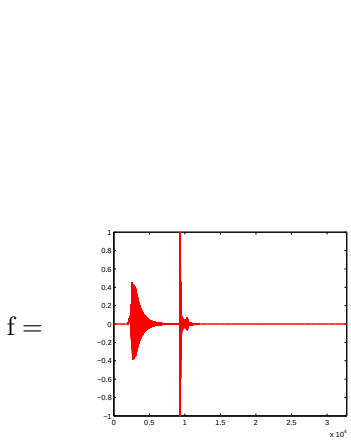
BIBO stability: $\ell^\infty \to \ell^\infty$ continuity is equivalent to $\mathrm{h} \in \ell^1$ if $\mathrm{Lf} = \mathrm{h} \star \mathrm{f}$.

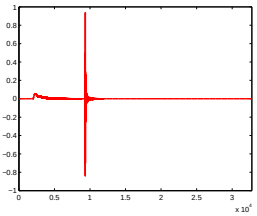
Causality: A LTI operator is causal if and only $\mathrm{Lf}[n]$ depends only on $\mathrm{f}[k]$ with $k \leq n$.

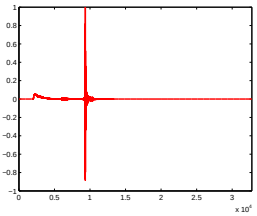
Proposition 2.1: If $\mathrm{Lf} = \mathrm{h} \star \mathrm{f}$ then L is causal if and only if $\mathrm{h}[k] = 0$ si $k < 0$.

Exponential: Exponential are eigenfunctions of L

$$\begin{aligned} \mathrm{L}\tau_k \mathrm{e}_s &= \tau_k \mathrm{L}\mathrm{e}_s \\ &= \mathrm{L}\mathrm{e}^{-ks}\mathrm{e}_s = \mathrm{e}^{-ks}\mathrm{L}\mathrm{e}_s \end{aligned}$$



L(f) =  Moving average

L(f) =  Low pass filter

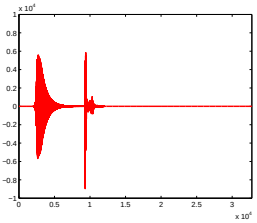
L(f) =  Differentiation

Figure 2.6: LTI examples

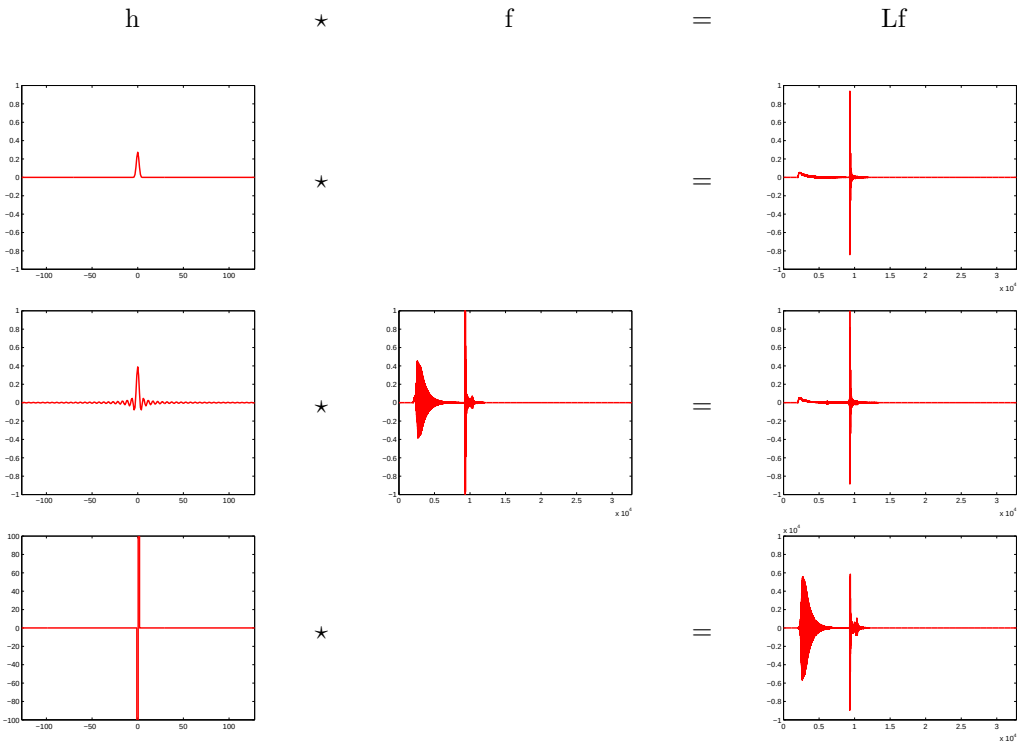


Figure 2.7: Impulse response

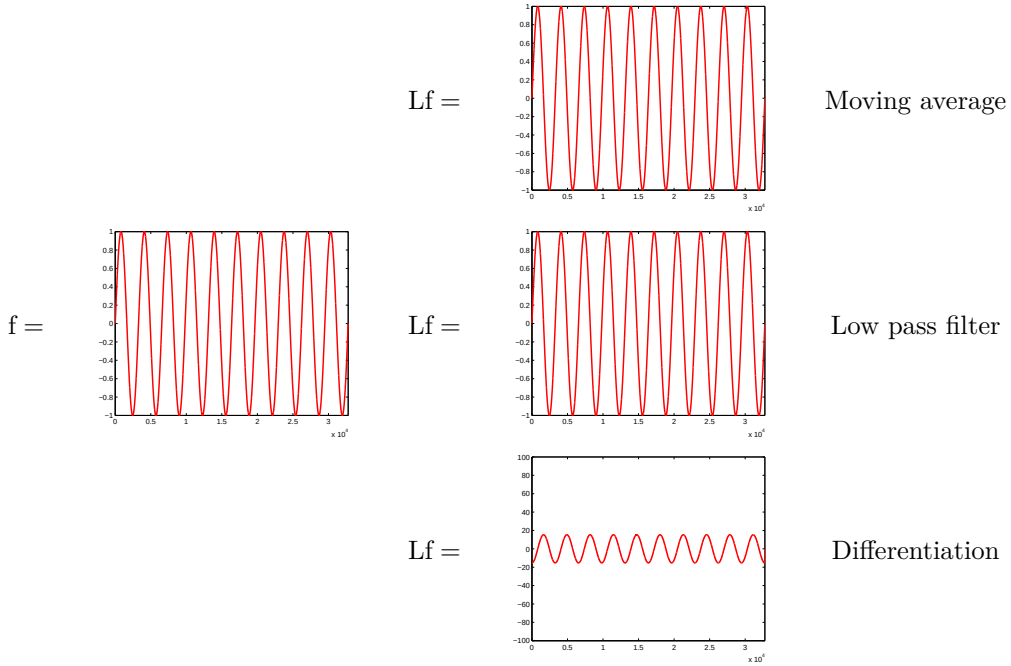


Figure 2.8: Exponential and LTI

which implies

$$\begin{aligned}\tau_k(\mathrm{e}_{-s}\mathrm{L}\mathrm{e}_s) &= \tau_k \mathrm{e}_{-s} \times \tau_k \mathrm{L}\mathrm{e}_s \\ &= e^{ks} e_{-s} \times e^{-ks} L e_s = e_{-s} L e_s\end{aligned}$$

i.e. $\mathrm{e}_{-s}\mathrm{L}\mathrm{e}_s$ is constant and thus $\mathrm{L}\mathrm{e}_s = \mathrm{H}(s)\mathrm{e}_s$ as in the analog case.

Note that as e_s is valued only at integer values $\mathrm{e}_s = \mathrm{e}_{s+i2\pi}$.

2.2.2 Distribution and Fourier transform of discrete signals

The simpℓ_1 case: Assume that $(\mathrm{f}[n])$ belongs to ℓ^1, we can define a *Fourier transform* of f by letting

$$\mathcal{F}\mathrm{f}(\omega) = \sum_{n\in\mathbb{Z}} \mathrm{f}[n] e^{-in\omega}$$

which is a continuous 2π-periodic function. To stress this periodicity, one often uses the notation $\widehat{\mathrm{f}}(e^{i\omega})$ instead of $\mathcal{F}\mathrm{f}(\omega)$. By construction,

$$\mathrm{f}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{\mathrm{f}}(e^{-i\omega}) e^{n\omega} \mathrm{d}\omega$$

that is an inverse Fourier transform.

We show now how to extend this construction to sequences that do not belong to ℓ^1 using the distribution theory.

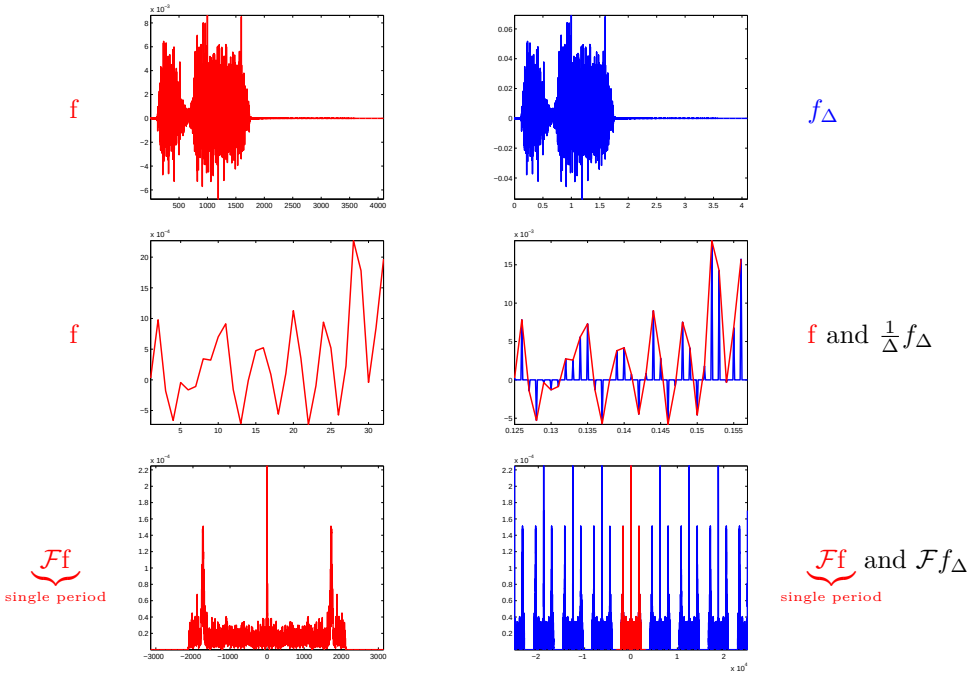


Figure 2.9: Discrete signal and periodic spectrum

Distribution representation: Let f[n] be a at most polynomially increasing sequence, in analogy with our discretization study, we can represent the sequence by the tempered distribution

$$f_\Delta = \Delta \sum_{n\in\mathbb{Z}} \mathrm{f}[n]\delta_{n\Delta}$$

where the f[n]s play the role of $f(n\Delta)$ of the discretization case. Δ is either naturally defined if indeed the sequence is the result of a discretization or can be chosen arbitrarily, with a classical choice of $\Delta = 1$.

Direct Fourier transform is defined through the Fourier transform of f_Δ:

$$\mathcal{F}_\Delta \mathrm{f} = \mathcal{F} f_\Delta = \Delta \sum_n f[n] e_{-in\Delta} = \Delta \mathcal{F}_1 \mathrm{f}(\cdot/\Delta)$$

which is a $2\pi/\Delta$ periodic (generalized) function in $\mathcal{S}'$ as soon as f grows at most polynomially. Note that, for $\Delta = 1$, this definition is consistant with the one for ℓ^1 sequences.

Inverse Fourier transform is defined through the tempered distribution theory. As $\mathcal{F}_\Delta$f is a $2\pi/\Delta$ periodic distribution and thus can be decomposed in Fourier series

$$\mathcal{F}_\Delta \mathrm{f} = \sum_{n\in\mathbb{Z}} \langle \mathcal{F}_\Delta \mathrm{f}, \frac{\Delta}{2\pi} \chi_{[-\pi/\Delta,\pi/\Delta]} e_{-in\Delta} \rangle e_{in\Delta}$$

and thus using $\mathcal{F}^{-1}e_{in\Delta} = \delta_{-n\Delta}$

$$\begin{aligned} f_\Delta &= \sum_{n\in\mathbb{Z}} \langle \mathcal{F}_\Delta \mathrm{f}, \frac{\Delta}{2\pi}\chi_{[-\pi/\Delta,\pi/\Delta]}e_{-in\Delta}\rangle \delta_{-n\Delta} \\ &= \sum_{n\in\mathbb{Z}} \frac{\Delta}{2\pi}\langle \mathcal{F}_\Delta \mathrm{f}, \chi_{[-\pi/\Delta,\pi/\Delta]}e_{in\Delta}\rangle \delta_{n\Delta}. \end{aligned}$$

This implies

$$\mathrm{f}[n] = \frac{1}{2\pi}\langle \mathcal{F}_\Delta \mathrm{f}, \chi_{[-\pi/\Delta,\pi/\Delta]}e_{in\Delta}\rangle$$

which becomes

$$\mathrm{f}[n] = \frac{1}{2\pi}\int_{-\pi/\Delta}^{\pi/\Delta} \mathcal{F}_\Delta \mathrm{f}(\omega)e^{in\Delta\omega}d\omega$$

as soon as $\mathcal{F}_\Delta \mathrm{f}$ belongs to L^1-loc, which is the case for instance if $\mathrm{f} \in \ell^1$

Link with Fourier series: $\Delta \mathrm{f}[-n]$ is nothing but the nth Fourier coefficient of the $2\pi/\Delta$ periodic (generalized) function $\mathcal{F}_\Delta \mathrm{f}$.

Proposition 2.2 (Parseval): If $\mathrm{f} \in \ell^2$ or equivalently $\mathcal{F}_\Delta \mathrm{f} \in L^2$-loc then

$$\Delta^2 \sum_{n\in\mathbb{Z}} |f[n]|^2 = \frac{\Delta}{2\pi}\int_{[-\pi/\Delta,\pi/\Delta]} \left|\mathcal{F}_\Delta \mathrm{f}(\omega)\right|^2 \mathrm{d}\omega$$

Duality:

- If f is discrete with step Δ then $\mathcal{F}f$ is periodic of period $2\pi/\Delta$
- If f is periodic of period T then $\mathcal{F}f$ is discrete with step $2\pi/T$

which is an instance of Fourier **global/local** duality (cf regularity properties).

DTFT (Discrete-Time Fourier Transform) corresponds to the choice $\Delta = 1$ and, denoting $\mathcal{F}\mathrm{f} = \mathcal{F}_1 \mathrm{f}$, leads to the formulas

$$\mathcal{F}\mathrm{f} = \sum_{n\in\mathbb{Z}} \mathrm{f}[n]e_{-in}$$

and

$$\mathrm{f}[n] = \frac{1}{2\pi}\langle \mathcal{F}\mathrm{f}, \chi_{[-\pi,\pi]}e_{in}\rangle$$

which can be rewritten if $\mathcal{F}\mathrm{f} \in L^1$-loc as

$$\mathrm{f}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{F}\mathrm{f}(\omega)e^{in\omega}d\omega$$

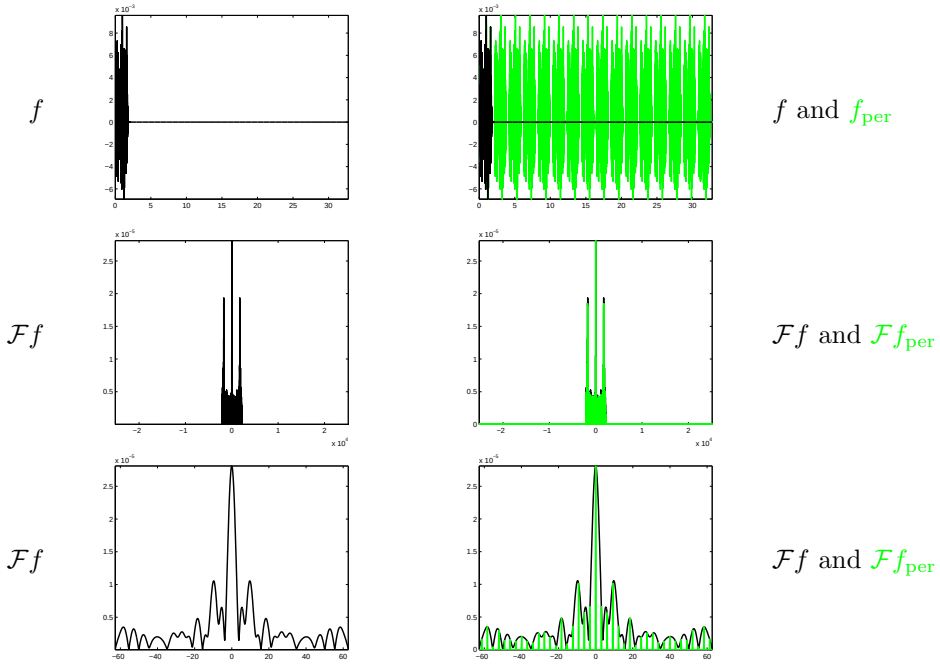


Figure 2.10: Periodic signal and discrete spectrum

More classical notation:

$$\widehat{\mathrm{f}}(e^{i\omega}) = \sum_{n\in\mathbb{Z}} \mathrm{f}[n]e^{-i\omega n} \quad \text{and} \quad \mathrm{f}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{\mathrm{f}}(e^{i\omega})e^{in\omega}d\omega$$

2.2.3 Convolution and transfer function

Proposition 2.3: $\mathcal{F}(\mathrm{h}\star\mathrm{g}) = \mathcal{F}\mathrm{h}\times\mathcal{F}\mathrm{f}$ provided the convolution is well defined and all the sequences h, f and $\mathrm{h}\star\mathrm{g}$ grow at most polynomially.

Transfer function: If $\mathrm{Lf} = \mathrm{h}\star\mathrm{f}$ then provided all the sequences involved grow at most polynomially

$$\begin{aligned}\mathcal{F}\mathrm{Lf} &= \mathcal{F}\mathrm{h}\times\mathcal{F}\mathrm{f}\\ \mathrm{Lf}[n] &= \langle \mathcal{F}\mathrm{h}\times\mathcal{F}\mathrm{f}, \chi_{[-\pi,\pi]}e_{in}\rangle\end{aligned}$$

which can be rewritten if $\mathcal{F}\mathrm{h}\times\mathcal{F}\mathrm{f}\in L^1$-loc as

$$\mathrm{Lf}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{F}\mathrm{h}(\omega)\mathcal{F}\mathrm{f}(\omega)e^{in\omega}d\omega.$$

Convolution corresponds to the multiplication by the transfer function $\mathcal{F}\mathrm{h}(\omega)$ in the Fourier domain.

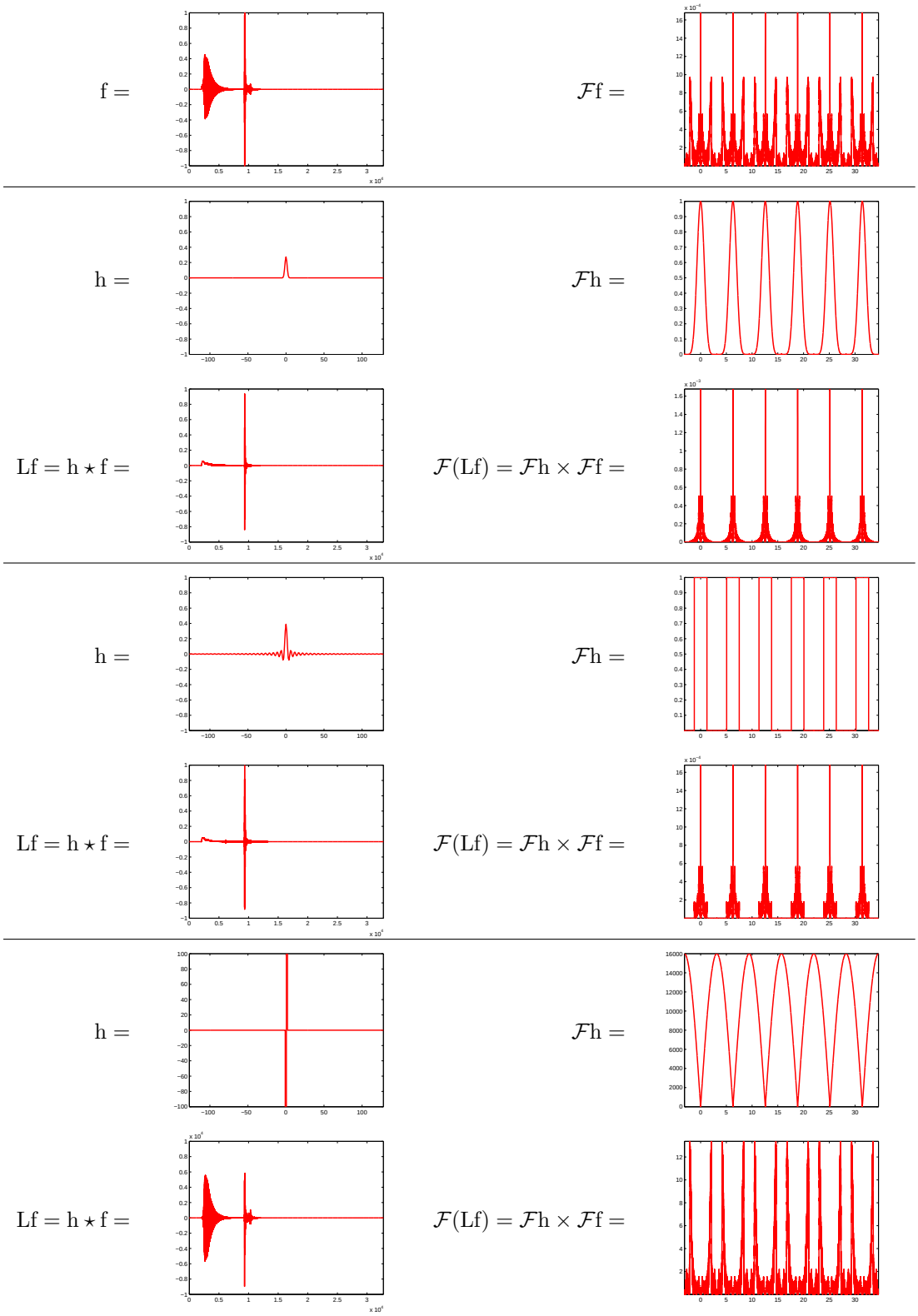


Figure 2.11: Frequential filtering

2.2.4 Filtering examples

Delay: $\mathrm{Lf}[n] = \mathrm{f}[n-\Delta]$

$$\mathrm{h}[n] = \delta_\Delta \quad \text{and} \quad \mathcal{F}\mathrm{h} = e_{-i\Delta}$$

Moving average: $\mathrm{Lf} = \frac{1}{2\Delta+1}\mathbf{1}_{[-\Delta,\Delta]} \star \mathrm{f}$

$$\mathrm{h} = \frac{1}{2\Delta+1}\mathbf{1}_{[-\Delta,\Delta]} \quad \text{and} \quad \mathcal{F}\mathrm{h} = \frac{\sin(\Delta+1/2)\omega}{(2\delta+1)\sin(\omega/2)}$$

Low pass filter: $\mathcal{F}\mathrm{Lf} = \mathcal{F}\mathrm{f} \times \sum_{n\in\mathbb{Z}} \mathbf{1}_{[-\omega_0,\omega_0]+n2\pi}$ $(0 < \omega_0 < \pi)$

$$\mathcal{F}\mathrm{h} = \sum_{n\in\mathbb{Z}} \mathbf{1}_{[-\omega_0,\omega_0]+n2\pi} \quad \text{and} \quad \mathrm{h}[n] = \omega_0 \operatorname{sinc} \omega_0 n$$

Derivation: $\mathrm{Lf}[n] = (\mathrm{f}[n+1] - \mathrm{f}[n-1])/2$

$$\mathrm{h} = (\delta_1 - \delta_{-1})/2 \quad \text{and} \quad \mathcal{F}\mathrm{h} = (e_{-i} - e_i)/2 = i\sin(\cdot)$$

Rational fraction and recursive filtering If $\mathcal{F}\mathrm{h}$ is **rational** in $e^{-i\omega}$:

$$\mathcal{F}\mathrm{h} = \frac{\sum_{k=0}^{K_a} a_k e^{-ik\omega}}{\sum_{k=0}^{K_b} b_k e^{-ik\omega}}$$

then, provided everything is well defined, $\mathrm{Lf} = \mathrm{h} \star \mathrm{f}$ and f

$$\mathcal{F}\mathrm{Lf} = \frac{\sum_{k=0}^{K_a} a_k e^{-ik\omega}}{\sum_{k=0}^{K_b} b_k e^{-ik\omega}} \mathcal{F}\mathrm{f}$$

Equivalently

$$\sum_{k=0}^{K_a} a_k e^{-ik\omega} \mathcal{F}\mathrm{Lf} = \sum_{k=0}^{K_b} b_k e^{-ik\omega} \mathcal{F}\mathrm{f} \Leftrightarrow \sum_{k=0}^{K_a} a_k \mathrm{Lf}[n-k] = \sum_{k=0}^{K_b} b_k \mathrm{f}[n-k]$$

We derive thus the following formula

$$a_0 \mathrm{Lf}[n] = -\sum_{k=1}^{K_a} a_k \mathrm{Lf}[n-k] + \sum_{k=0}^{K_b} b_k \mathrm{f}[n-k]$$

which corresponds to a recursive implementation of the filter. Note that there is an initialization issue as the past of Lf should be known...

If the filter is causal and $\mathrm{f}[n]$ is assumed to be 0 for $n < 0$ then $\mathrm{Lf}[n] = 0$ for $n < 0$ and thus this implementation can be used...

2.2.5 z-transform to analyze causality and stability

Extension of Fourier analysis: From the unit circle

$$\mathcal{F}\mathrm{f} : \omega \mapsto \sum_{n\in\mathbf{Z}} \mathrm{f}[n]e^{-in\omega} \quad (= \widehat{\mathrm{f}}(e^{iw}))$$

to the complex plane

$$\mathfrak{F}\mathrm{f} : z \mapsto \sum_{n\in\mathbf{Z}} \mathrm{f}[n]z^{-n} \quad (= \widehat{\mathrm{f}}(z)).$$

This extension shares similarity with the Laplace transform.

Laurent series (at 0) are expressions of the form

$$\sum_{n\in\mathbf{Z}} \mathrm{f}[n]z^{-n}.$$

Their domain of convergence is a possibly empty annulus (ring).

Inversion formula through analytic function theory:

$$\mathrm{f}[n] = \frac{1}{2\pi i}\int_\gamma \frac{\mathfrak{F}\mathrm{f}(z)}{z^{n+1}}\mathrm{d}z$$

with γ is counterclockwise around a closed, rectifiable path containing no self-intersections, enclosing 0 and lying in the annulus of convergence.

Classical examples:

$$\begin{aligned}
\mathfrak{F}\delta_\Delta &= z^{-\Delta} && (z\in\mathbb{C})\\
\mathfrak{F}a^n\mathrm{h} &= \mathfrak{F}\mathrm{h}(\cdot/a)\\
\mathfrak{F}n\mathrm{h} &= -z(\mathfrak{F}\mathrm{h})'\\
\mathfrak{F}(\mathbf{1}_{n\geq 0}) &= \frac{1}{1-z^{-1}} && (|z|>|1|)\\
\mathfrak{F}(-\mathbf{1}_{n<0}) &= \frac{1}{1-z^{-1}} && (|z|<|1|)\\
\mathfrak{F}(a^n\mathbf{1}_{n\geq 0}) &= \frac{1}{1-az^{-1}} && (|z|>|a|)\\
\mathfrak{F}(-a^n\mathbf{1}_{n<0}) &= \frac{1}{1-az^{-1}} && (|z|<|a|)\\
\mathfrak{F}(n\mathbf{1}_{n\geq 0}) &= \frac{z^{-1}}{(1-z^{-1})^2} && (|z|>|1|)\\
\mathfrak{F}(n^2\mathbf{1}_{n\geq 0}) &= \frac{z^{-1}(1+z^{-1})}{(1-z^{-1})^3} && (|z|>|1|)
\end{aligned}$$

Proposition 2.4: A filter is stable ($\in \ell^1$) iff the unit circle lies in its dom. of conv.

Recursive filtering and pole: A recursive filter h has a z-transform $\mathfrak{F}h(z) = P(z^{-1})/Q(z^{-1})$ where we assume that P and Q have no common roots. If it is stable then Q has no root of modulus 1 and it is furthermore causal then Q has no root of modulus < 1. For $\mathfrak{F}h(z)$ this correspond to having no pole of modulus > 1.

2.3 Finite Signal Processing

2.3.1 Finite Signal and periodization

Finite (Discrete) signal: is a finite sequence $\{\mathrm{f}[n]\}_{0 \leq n < N}$.

Boundary issues: What is going on before 0 and after $N-1$?

- Nothing (zero-padding): $\mathrm{f}[n] = 0$ if $n < 0$ or $n \geq N$,
- The same thing (periodization): $\mathrm{f}[n] = \mathrm{f}[n \mod N]$,
- Almost the same thing (symmetrization and periodization): different choices for the symmetrization operator...

Periodization and distributions:

$$f = \sum_n \mathrm{f}[n \mod N]\delta_{n\Delta}$$

is a discrete (of step Δ) and periodic (of period $N\Delta$) distribution whose Fourier transform is thus a periodic (of period $2\pi/\Delta$) and discrete (of step $2\pi/(N\Delta)$ distribution

$$\mathcal{F}f = \sum_n \widehat{\mathrm{f}}_\Delta[n \mod N]\delta_{2\pi/(N\Delta)}$$

where a simple computation (using Poisson formula) shows that

$$\widehat{\mathrm{f}}_\Delta[k] = \sum_{n=0}^{N-1} \mathrm{f}[n]e^{-i2\pi kn/N}.$$

which is thus specified by N values, i.e. a finite signal.

Finite Fourier Transform a.k.a. Discrete Fourier Transform is defined in $\mathbb{C}^N$ by

$$\widehat{\mathrm{f}}[k] = \sum_{n=0}^{N-1} \mathrm{f}[n]e^{-i2\pi kn/N}$$

whose inverse is given by

$$\mathrm{f}[n] = \frac{1}{N}\sum_{n=0}^{N} \mathrm{f}[k]e^{i2\pi kn/N}.$$

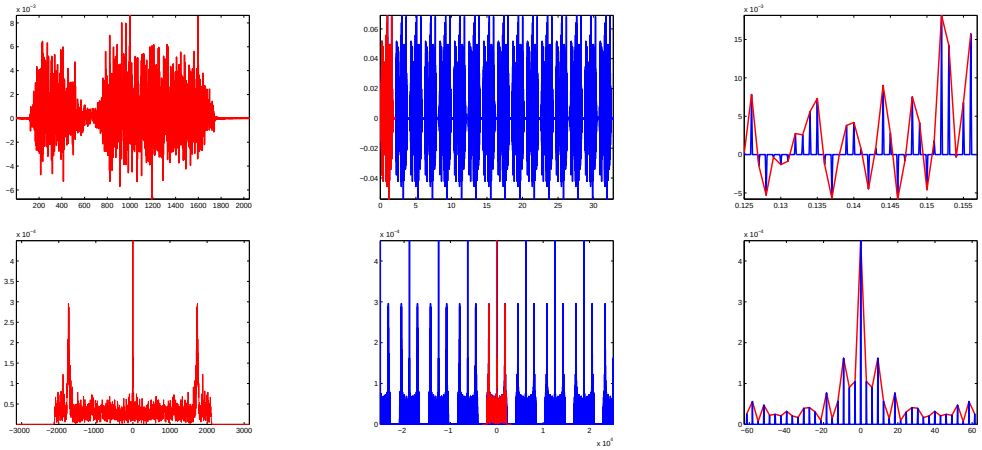


Figure 2.12: DFT

This could be reinterpreted in term of the orthonormality in $\mathbb{C}^N$ of

$$\{\frac{1}{\sqrt{N}}e^{i2\pi k\cdot/N}\}_{0\leq k<N}$$

but fails to convey all the subtleties of the DFT...

2.3.2 LTI, Circular convolution and Fourier filtering

LTI: Linear and Translation Invariant for finite signals means Linear and Translation Invariant for periodized signals.

Convolution: If $\mathrm{Lf} = \mathrm{h} \star \mathrm{f}$ then

$$\begin{aligned}
\mathrm{Lf}[n] &= \sum_{k\in\mathbb{Z}} \mathrm{h}[k]\mathrm{h}[n-k] \\
&= \sum_{k\in\mathbb{Z}} \mathrm{h}[k]\mathrm{h}[n-k \mod N] \\
&= \sum_{k=0}^{N-1} \left(\sum_{l\in\mathbb{Z}} \mathrm{h}[k+lN] \right) \mathrm{h}[n-k \mod N] \\
&= \sum_{k=0}^{N-1} \widetilde{\mathrm{h}}[k]\mathrm{h}[n-k \mod N]
\end{aligned}$$

where $\widetilde{\mathrm{h}}$ is a periodized version of h

$$\widetilde{\mathrm{h}}[k] = \sum_{l\in\mathbb{Z}} \mathrm{h}[k+lN].$$

This *convolution* is often called circular convolution and is denoted $\circledast$.

Proposition 2.5 (Fourier and convolution):

$$\widehat{\mathrm{h} \circledast \mathrm{f}} = \widehat{\mathrm{h}}\widehat{\mathrm{f}}$$

2.3.3 FFT

Naive implementation: In order to compute

$$\widehat{\mathrm{f}}[k] = \sum_{n=0}^{N-1} \mathrm{f}[n] e^{-i2\pi kn/N}$$

for $0 \leq k < N$, the easiest way is to compute those N scalar products, each requiring N multiplications and $N-1$ additions. The total number of operations is thus N^2 multiplications and $N^2 - N$ additions. As soon as N is large, this may prevent the use of such a transform in *real life* situation.

The even N trick: If N is even then

$$\widehat{\mathrm{f}}[2k'] = \sum_{n=0}^{N/2-1} (\mathrm{f}[n] + \mathrm{f}[N/2+n])\, e^{-i2\pi k' n/(N/2)}$$
$$\widehat{\mathrm{f}}[2k'+1] = \sum_{n=0}^{N/2-1} (\mathrm{f}[n] - \mathrm{f}[N/2+n])\, e^{-i2\pi/N} e^{-i2\pi k' n/(N/2)}$$

and thus the DFT of size N can be computed from 2 DFT of 2 signals of size $N/2$ obtained by N additions and $N/2$ multiplications. Using the naive previous implementation for those DFT yields thus $N/2 + 2(N/2)^2$ multiplications and $2(N/2)^2$ additions. There is thus a gain of almost a factor 2!

The $N = 2^p$ case: In this case, this idea can be reused p times (the DFT of a signal of size 1 being trivial). Analyzing its complexity yields

$$\text{Nb mult}(2^p) = 2\text{Nb mult}(2^{p-1}) + 2^p/2$$
$$\text{Nb add}(2^p) = 2\text{Nb add}(2^{p-1}) + 2^p$$

and thus

$$\frac{\text{Nb mult}(2^p)}{2^p} = \frac{\text{Nb mult}(2^{p-1})}{2^{p-1}} + 1/2$$
$$\frac{\text{Nb add}(2^p)}{2^p} = \frac{\text{Nb add}(2^{p-1})}{2^{p-1}} + 1$$

so that

$$\text{Nb mult}(2^p) = p/2 \times 2^p$$
$$\text{Nb add}(2^p) = p \times 2^p.$$

In term of N, one obtains thus a complexity of order $N \log N$ which is much smaller than N^2!

FFT (Fast Fourier Transform): This algorithm has been proposed in 1965 by Cookey and Tuckey when $N = 2^p$ and is a crucial step in Discrete Signal Processing. Variations are possible when $N \neq 2^p$ using factorization in prime factors.

2.3.4 What is really computing the FFT algorithm?

Naive interpretation: It computes the Fourier transform of a function f...

Not so naive interpretation: It computes the Fourier transform of

$$f_{\Delta,\mathrm{per}} = \Delta \sum_n f((n \mod N)\Delta)\delta_{n\Delta}$$

or more accurately the N-periodic coefficients of the Dirac comb

$$\mathcal{F}f_{\Delta,\mathrm{per}} = \sum_n \widehat{\mathrm{f}}[n]\delta_{n2\pi/(N\Delta)}.$$

Relationship between the two transforms: As

$$f_{\Delta,\mathrm{per}} = \left((f \times \chi_{[0,N\Delta)}) \times \sum_n \delta_{n\Delta} \right) \star \Delta \sum_k \delta_{kN\Delta}$$

and thus at least formally

$$\mathcal{F}f_{\Delta,\mathrm{per}} = 1/2\pi \left((\mathcal{F}f \star \mathcal{F}(\chi_{[0,N\Delta)})) \star 2\pi \sum_n \delta_{n2\pi/\Delta} \right) \times \sum_k \delta_{2\pi/(N\Delta)}$$

which implies

$$\widehat{\mathrm{f}}[n] = \sum_{k\in\mathbb{Z}} (\mathcal{F}f \star \mathcal{F}(\chi_{[0,N\Delta)}))(2\pi(n+kN)/(N\Delta)).$$

We compute thus the discretized version of the periodization of the convolution of the Fourier transform of f with a sinc function!

Remark: This analysis holds only if $\mathcal{F}f$ is compactly supported. Imagine for instance that f is equal to 0 in $[0, N]$, then there is no chance to recover something different than 0 from the samples in $[0, N]$... but if $\mathcal{F}f$ is compactly supported then f is analytic and thus equal to 0 if it vanishes on an interval.

Windowing: As the sinc decays slowly, the convolution may yields a significant modification of the Fourier transform. To mitigate this effect, one can replace the multiplication by $\mathbf{1}_{[0,N\Delta)]}$ by the multiplication by a smoother function $W_{[0,N\Delta]}$ which is equal to 1 in the middle of the interval and decays to 0 at the boundaries.

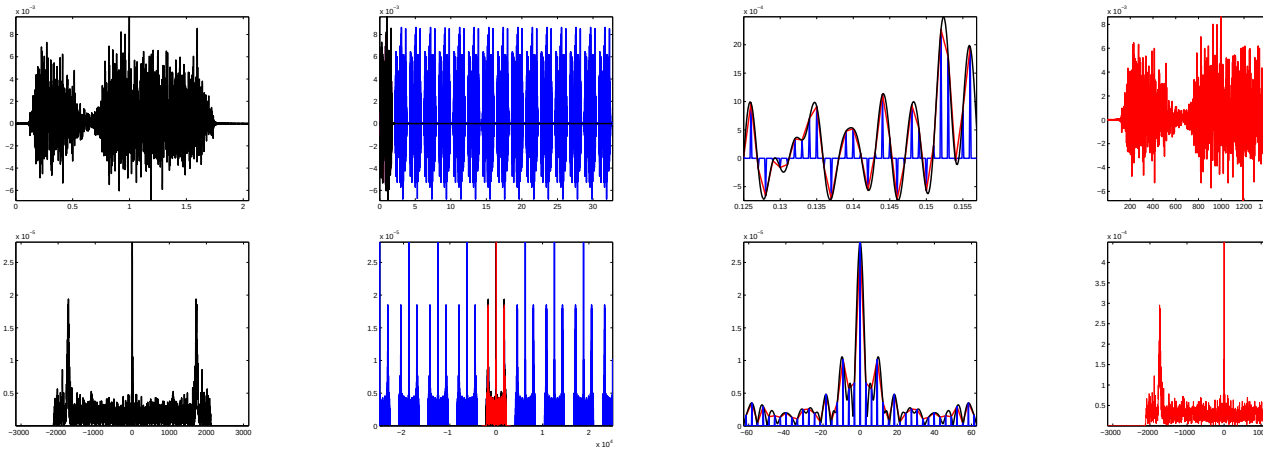


Figure 2.13: From Fourier to FFT

Part II

Voice processing: Time-Frequency Analysis and Stationary process modeling

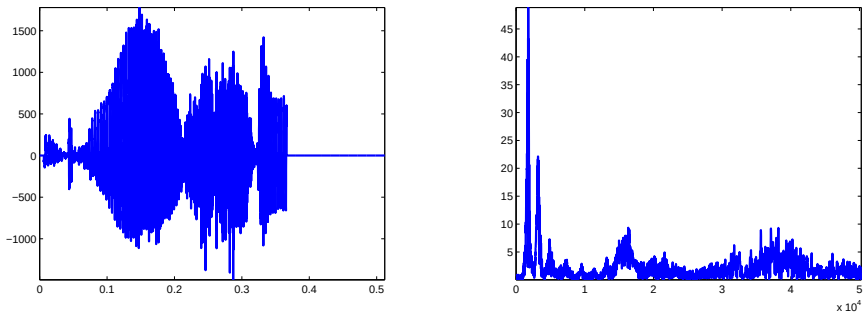
Chapter 3

Time-Frequency Analysis

In the previous part, we have studied the Fourier transform of a signal which gives a frequency view of it. In particular, it is hard to read the time behavior which was obvious is the time representation. We should show here there is an intermediate way in which one can analyze the signal simultaneously in time and frequency.

Assume one wants to analyze the sound of a women saying *Greasy*. So far we have two views of the correspond signal: the temporal one and the spectral one using its Fourier transform.

Although those representation convey all the information cotained in the signal, they do not correspond to the intuition that a sound is made of notes having a certain frequency and a certain location, as they are described in a musical score. In this chapter, we will show how to obtain a *score* in the time-frequency plane.



f $\quad$ $\widehat{f}$

$\omega \geq 0$

Figure 3.1: Spatial or spectral view of a signal

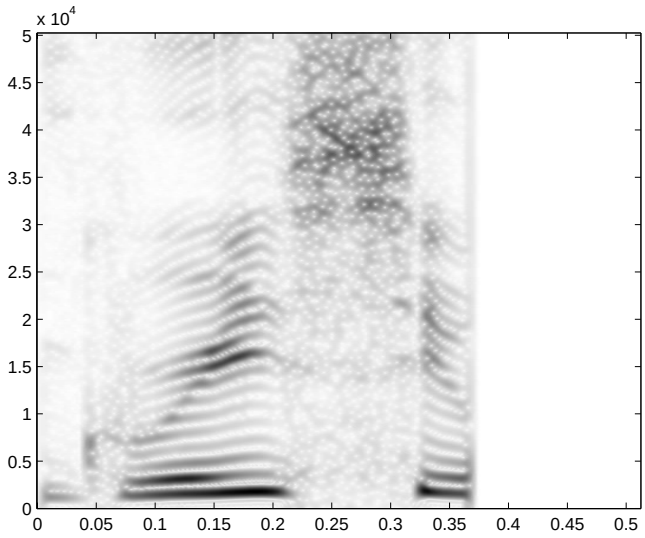


Figure 3.2: Time-Frequency score

3.1 Time frequency atoms and Windowed Fourier Transform

3.1.1 Localized Fourier transform

The Fourier transform and localization: When f belongs to L^1, one can write

$$\widehat{f}(\omega) = \int_{\mathbb{R}} f(t)e^{-i\omega t}\mathrm{d}t.$$

If one wants to study the behavior of f around 0 a natural idea is to multiply f by a bounded real window w equal to 1 around 0 and vanishing away from 0 and to compute

$$\int_{\mathbb{R}} f(t)w(t)e^{-i\omega t}\mathrm{d}t.$$

Windows: The most simple example of such a window is the characteristic set of $[-h, h]$. As we have seen in the previous part, if f is band-limited, the resulting transform is the Fourier transform of f convolved by a sinc function.

We will use more regular window w and define for $f \in L^1$ its windowed Fourier transform at 0 by

$$Sf(0,\omega) = \int_{\mathbb{R}} f(t)w(t)e^{-i\omega t}\mathrm{d}t$$

and more generally its windowed Fourier transform at u by

$$Sf(u,\omega) = \int_{\mathbb{R}} f(t)w(t-u)e^{-i\omega t}\mathrm{d}t.$$

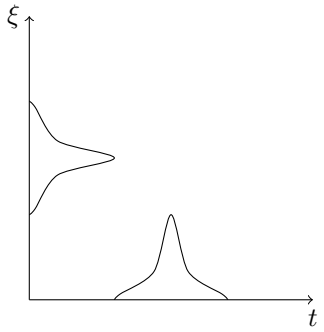


Figure 3.3: Time-Frequency atom

The L^2case: If we assume that w belongs to L^2 then for any $f \in L^2$

$$Sf(u,\omega) = \int_{\mathbb{R}} f(t)w(t-u)e^{-i\omega t}\mathrm{d}t$$

is well defined and equal to

$$= \langle f, w(\cdot - u)e_{i\omega}\rangle$$

where the scalar product is now the hermitian scalar product.

Time-Frequency atoms: From now on, we assume that $f \in L^2$ and $w \in L^2$. We do not assume anymore that w is real valued and define

$$Sf(u,\omega) = \int_{\mathbb{R}} f(t)\overline{w(t-u)}e^{-i\omega t}\mathrm{d}t$$

that the windowed Fourier transform can be interpreted as the collection of all scalar products of f with the probes $w_{u,\omega} = w(\cdot - u)e_{i\omega}$ which are called time-frequency atoms.

3.1.2 Time-frequency atoms position

Fourier domain probes: Using Parseval equality, we deduce immediately that

$$\begin{aligned} Sf(u,\omega) &= \langle f, w_{u,\omega}\rangle \\ &= \frac{1}{2\pi}\langle \widehat{f}, \widehat{w_{u,\omega}}\rangle \end{aligned}$$

where $\widehat{f}$ denotes $\mathcal{F}f$ and thus the Windowed Fourier transform can also be interpreted as the collection of all scalar products of $\widehat{f}$ with the probes $\frac{1}{2\pi}\widehat{w_{u,\omega}}$.

Atoms centers: As $w_{u,\omega} = \overline{w(\cdot - u)}e^{i\omega t}$ and thus $\widehat{w_{u,\omega}}(\xi) = e^{iu\xi}\widehat{w}(\cdot - \omega)$, if we assume than w is *centered* on 0 and $\widehat{w}$ is centered on 0 then $w_{u,\omega}$ is *centered* on u and $\widehat{w_{u,\omega}}$ is centered on ω.

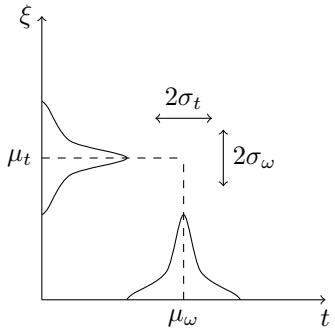


Figure 3.4: Time-Frequency atom localization

Spatial and spectral densities and means: More precisely, to any $g \in L^2 \setminus \{0\}$, we can associate its spatial density $|g|^2/\|g\|_2^2$ and its spectral density $|\widehat{g}|^2/\|\widehat{g}\|_2^2$. We define then its spatial mean by

$$\mu_t(g) = \int t \frac{|g(t)|^2}{\|g\|_2^2} \mathrm{d}t$$

and its spectral mean by

$$\mu_\omega(g) = \int \omega \frac{|\widehat{g}(\omega)|^2}{\|\widehat{g}(\omega)\|_2^2} \mathrm{d}\omega.$$

Remark that those definitions are coherent with the idea that if g is symmetric around 0 then its mean should be 0.

Atoms means: If we apply those definitions to $w_{u,\omega}$ one obtains that

$$\mu_t(w_{u,\omega}) = \mu_t(w_{u,\omega'}) = u + \mu_t(w) \quad \text{and} \quad \mu_\omega(w_{u,\omega}) = \mu_\omega(w_{u',\omega}) = \omega + \mu_\omega(w).$$

If we impose that $\mu_t(w) = \mu_\omega(w) = 0$, that is w is centered spatially and spectrally then

$$\mu_t(w_{u,\omega}) = \mu_t(w_{u,\omega'}) = u \quad \text{and} \quad \mu_\omega(w_{u,\omega}) = \mu_\omega(w_{u',\omega}) = \omega$$

3.1.3 Time-frequency atoms localization

3.1.4 Densities and dispersion

We can also measure the dispersion of a function around its spatial and spectral mean by defining its spatial standard deviation σ_t and its spectral standard deviation σ_ω:

$$\sigma_t^2(g) = \int (t - \mu_t(g))^2 \frac{|g(t)|^2}{\|g\|_2^2} \mathrm{d}t \quad \text{and} \quad \sigma_\omega^2(g) = \int (\omega - \mu_\omega(g))^2 \frac{|\widehat{g}(\omega)|^2}{\|\widehat{g}\|_2^2} \mathrm{d}t$$

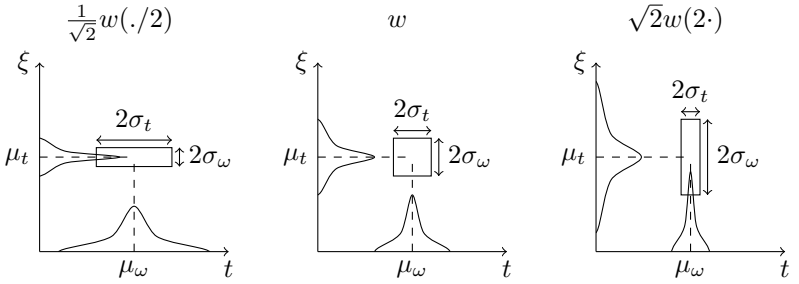


Figure 3.5: Heisenber boxes

3.1.5 Atoms localization:

If we apply those definitions to $w_{u,\omega}$ one obtains that

$$\sigma_t(w_{u,\omega}) = \sigma_t(w) \quad \text{and} \quad \sigma_\omega(w_{u,\omega}) = \sigma_\omega(w)$$

so that all atoms have the same localization properties.

3.1.6 Heisenberg incertitude theorem

Heisenberg boxes: In a Time-Frequency plane, each atom $w_{u,\omega}$ can be represented by a Heisenberg box of size $2\sigma_t(w) \times 2\sigma_\omega(w)$ centered on (u, ω). For a given w, all those boxes have the same size. Intuitively, the smaller the box the better.

Heinseberg box and scaling: A first attempt is to replace w by $w(./s)$, and thus $\hat{w}$ by $s\hat{w}(s\cdot)$, a straightforward computation shows that

$$\mu_t\left(w(\cdot/s)\right) = \mu_t(w)/s \quad \text{and} \quad \mu_\omega\left(w(\cdot/s)\right) = s\mu_\omega(w)$$

while

$$\sigma_t\left(w(\cdot/s)\right) = \sigma_t(w)/\sqrt{s} \quad \text{and} \quad \sigma_\omega\left(w(\cdot/s)\right) = \sqrt{s}\sigma_\omega(w)$$

So that any gain is one domain is exactly compensated by a loss in the other one.

Heisenberg box and w: The only solution seems to be a clever choice of w. Unfortunately, there is a lower limit on the product $\sigma_t \times \sigma_\omega$.

Theorem 3.1: If $w \in L^2 \setminus \{0\}$ then

$$\sigma_t^2(g)\sigma_w^2(g) \geq \frac{1}{4}$$

with equality if and only if there exist $a \in \mathbb{C}$, $b > 0$, $(u, \omega) \in \mathbb{R}^2$ such that

$$g(t) = ae^{-b((t-u)^2}e^{i\omega t}$$

Proof: Remark that $\sigma_t(g)$ is equal to 0 if and only if g is supported on a singleton, which is impossible if $g \in L^2 \setminus \{0\}$. Now if $\sigma_t(g) = +\infty$ or $\sigma_\omega(g) = +\infty$ the result is trivial and thus we assume from now on that $\sigma_t(g) < +\infty$ or $\sigma_\omega(g) < +\infty$.

Without loss of generality, we may assume further that

$$\mu_t(g) = 0 \quad \text{and} \quad \mu_\omega(g) = 0.$$

Indeed we have seen that for (u, ω), $\sigma_t(g_{u,\omega}) = \sigma_t(g)$ and $\sigma_\omega(g_{u,\omega}) = \sigma_\omega(g)$ so we can replace g by $g_{u,\omega}$ with $u = -\mu_t(g)$ and $\omega = -\mu_\omega(g)$ which satisfies those assumptions.

Assume for the moment that $\sigma_t(g) < +\infty$ and $\sigma_\omega(g) < +\infty$ implies g continuous, $\lim_{|t|\to\infty} t|g(t)|^2 = 0$ and $g' \in L^2$

We have thus

$$\begin{aligned}\sigma_t^2(g)\sigma_w^2(g) &= \int t^2 \frac{|g(t)|^2}{\|g\|_2^2}\mathrm{d}t \times \int \omega^2 \frac{|\widehat{g}(\omega)|^2}{\|\widehat{g}\|_2^2}\mathrm{d}\omega \\ &= \frac{1}{\|g\|_2^4}\int t^2|g(t)|^2\mathrm{d}t \times \frac{1}{2\pi}\int |i\omega\widehat{g}(\omega)|^2\mathrm{d}\omega\end{aligned}$$

as $-i\omega\widehat{g} = \widehat{g'}$, using Parseval equality we deduce

$$= \frac{1}{\|g\|_2^4}\int t^2|g(t)|^2\mathrm{d}t \times \int |g'(t)|^2\mathrm{d}t$$

By Cauchy-Schwarz, we obtain

$$\geq \frac{1}{\|g\|_2^4}\left(\frac{1}{2}\left|\int tg(t)\bar{(}g'(t))dt\right| + \frac{1}{2}\left|\int t\overline{g(t)}g'(t)dt\right|\right)^2$$

By triangular inequality

$$\geq \frac{1}{4\|g\|_2^4}\left|\int t\left(g(t)\bar{(}g'(t)) + \overline{g(t)}g'(t)\right)dt\right|^2$$

Let L be this quantity

$$\begin{aligned}L &= \frac{1}{4\|g\|_2^4}\left|\int t\left(g(t)\bar{(}g'(t)) + \overline{g(t)}g'(t)\right)dt\right|^2 \\ &= \frac{1}{4\|g\|_2^4}\left|\int t(|g(t)|')^2\mathrm{d}t\right|^2\end{aligned}$$

as g is continuous a by part integration yields

$$= \frac{1}{4\|g\|_2^4}\left|\left[t|g(t)|^2\right]_\infty^{+\infty} - \int |g(t)|^2\mathrm{d}t\right|^2$$

and as $\lim_{|t|\to\infty} t|g(t)|^2 = 0$

$$L = \frac{1}{4}.$$

The Cauchy-Schwarz inequality is an equality if it exists $(\lambda, \gamma) \in \mathbb{C}^2$ such that

$$\lambda t g(t) = g'(t), \quad \gamma t \overline{g(t)} = \overline{g'(t)}.$$

We have thus here $\lambda = \overline{\gamma}$. Now the triangular inequality is an equality if it exists $\nu \geq 0$ such that

$$\int t g(t) \overline{g'(t)} dt = \nu \int t \overline{g(t)} g'(t) dt$$
$$\overline{\lambda} \int |t|^2 |g(t)|^2 \mathrm{d}t = \nu\lambda \int |t|^g(t)^2 |g(t)|^2 \mathrm{d}t$$

which implies $\lambda \in \mathbb{R}$. Now $\lambda t g(t) = g'(t)$ implies $g(t) = a e^{\lambda t^2/2}$ with $a \in \mathbb{C}$. Let $b = -\lambda/2$, as $g \in L^2$, we have $b = -\lambda/2 > 0$ which concludes the proof as every function $g = a e^{-b(t-u)^2} e^{ei\omega t}$ satisfies the equality constraints.

We should now go back to the proof that $\sigma_t(g) < +\infty$ and $\sigma_\omega(g) < +\infty$ implies g continuous, $\lim_{|t|\to\infty} t|g(t)|^2 = 0$ and $g' \in L^2$.

As $g \in L^2$, $g \in \mathcal{S}'$ and thus $g' \in \mathcal{S}'$. Now $\mathcal{F}g' = i\omega \mathcal{F}g = i\omega \widehat{g}$ and thus, as $\sigma_\omega^2(g) = \int \omega^2 \frac{|\widehat{g}(\omega)|^2}{\|\widehat{g}\|^2} \mathrm{d}\omega$, $\sigma_\omega(g) < \infty$ implies $\mathcal{F}g' \in L^2$ which in turns implies $g' \in L^2$.

The continuity of g is more involved. Let ϕ_n be a sequence of function in $\mathcal{S}$ such that $\|g - \phi_n\|_2^2 + \|g' - \phi_n'\|_2^2$ converge to 0. The sequence ϕ_n is thus a Cauchy sequence for the norm $\phi \mapsto \|\phi\|_2^2 + \|\phi'\|_2^2$. Now for any $\phi \in \mathcal{S}$,

$$\begin{aligned}
|\phi(t)| &= \left| \frac{1}{2\pi} \int \widehat{\phi}(\omega) e^{i\omega t} \mathrm{d}\omega \right| \\
&\leq \frac{1}{2\pi} \int \left| \widehat{\phi}(\omega) \right| \mathrm{d}\omega \\
&\leq \frac{1}{2\pi} \left(\int \frac{1}{1+\omega^2} \mathrm{d}\omega \right)^{1/2} \left(\int (1+\omega^2) |\widehat{\phi}(\omega)|^2 \mathrm{d}\omega \right)^{1/2} \\
&\leq \frac{1}{\sqrt{2\pi}} \left(\int \frac{1}{1+\omega^2} \mathrm{d}\omega \right)^{1/2} \left(\|\phi\|_2^2 + \|\phi'\|_2^2 \right)^{1/2}.
\end{aligned}$$

Thus ϕ_n is also a Cauchy sequence for the norm $\|\cdot\|_\infty$ which implies that ϕ_n converges toward a continuous function and thus by unicity of the limit g is continuous.

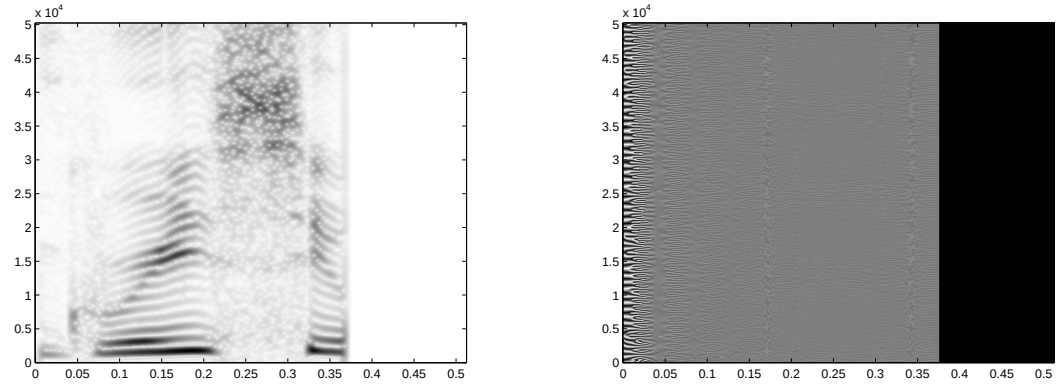
Now for any $m \leq M$,

$$\int_m^M t(|g(t)|^2)' \mathrm{d}t = \left[t|g(t)|^2 \right]_m^M - \int_m^M |g(t)|^2 \mathrm{d}t$$

and, as both integral have a finite limit when either m or M goes to respectively $-\infty$ and $+\infty$, so has $t|g(t)|^2$. Let l_+ be the limit at $+\infty$, $|g(t)|^2$ is equivalent to l_+/t and thus, as $g \in L^2$, l_+ cannot be anything but 0. A similar reasoning for $-\infty$ concludes the proof.

Examples:

- $w = \mathbf{1}_{[-a,a]}$: $\mu_t(w) = \mu_\omega(w) = 0$, $\sigma_t = a/\sqrt{3}$ and $\sigma_\omega = +\infty$,
- $w = e^{-bt^2}$: $\mu_t(w) = \mu_\omega(w) = 0$, $\sigma_t = \sqrt{2}/\sqrt{b}$ and $\sigma_\omega = \sqrt{b}/(2\sqrt{2})$.



$|Sf(u,\omega)|$ $\quad$ $\operatorname{Arg} Sf(u,\omega)$

Figure 3.6: STFT representation of Greasy

3.2 Windowed Fourier transform or STFT

3.2.1 Definition

To simplify the computations, we assume from now on that $\|w\|_2^2 = 1$.

The Windowed Fourier transform, or Short Time Fourier Transform , of $f \in L^2$ is then defined as

$$(u,\omega) \mapsto Sf(u,\omega) = \int f(t)\overline{w(t-u)}e^{-i\omega t}\mathrm{d}t = \langle f, w_{u,\omega}\rangle.$$

As we are working in a L^2 setting, we define the Spectrogram of f which measure the *local energy* of f as

$$(u,\omega) \mapsto |Sf(u,\omega)|^2.$$

3.2.2 Completeness and stability of the STFT representation

Theorem 3.2: If $w \in L^2 \cap L^1$ and $\|w\|_2 = 1$, then for any $f \in L^2$:

$$f(t) \overset{L^2}{=} \int\int Sf(u,\omega)w(t-u)e^{i\omega t}\mathrm{d}u\mathrm{d}\omega$$
$$\int |f(t)|^2\mathrm{d}t = \frac{1}{2\pi}\int\int |Sf(u,\omega)|^2\,\mathrm{d}u\mathrm{d}\omega$$

Note that this is similar to an orthogonal basis decomposition as it can be rewritten as

$$f(t) \overset{L^2}{=} \int\int \langle f, w_{u,\omega}\rangle w_{u,\omega}\mathrm{d}u\mathrm{d}\omega$$
$$\int |f(t)|^2\mathrm{d}t = \frac{1}{2\pi}\int\int |\langle f, w_{u,\omega}\rangle|^2\,\mathrm{d}u\mathrm{d}\omega$$

but with a lot of redundancy!

Lapped transform interpretation: Assume $f \times \overline{w}(\cdot - u) \in L^2$ so that $Sf(u,\omega) = \widehat{f \times \overline{w}}(\omega) \in L^2$. Assume further that $\omega \mapsto Sf(u,\omega)$ belongs to L^1, then

$$f(t)\overline{w(t-u)} = \frac{1}{2\pi}\int Sf_{u,\omega}e^{i\omega t}\mathrm{d}\omega$$

and thus multiplying by $w(t-u)$

$$f(t)\left|w(t-u)\right|^2 = \frac{1}{2\pi}\int Sf_{u,\omega}w(t-u)e^{i\omega t}\mathrm{d}\omega$$

while

$$\int |f(t)|^2|\overline{w(t-u)}|^2\mathrm{d}t = \frac{1}{2\pi}\int |Sf_{u,\omega}|^2\mathrm{d}\omega$$

and thus

$$\int |f(t)|^2|w(t-u)|^2\mathrm{d}t = \frac{1}{2\pi}\int |Sf_{u,\omega}|^2\mathrm{d}\omega.$$

Integrating those equalities along u yields the result as $\int |w(t-u)|^2\mathrm{d}u = 1$.

Discretization of u: Note that following this (formal) analysis hints that one could discretize the position u and obtains

$$f(t) \overset{L^2}{=} \sum_{u\in\Gamma_u}\int \langle f, w_{u,\omega}\rangle w_{u,\omega}\mathrm{d}u\mathrm{d}\omega$$

$$\int |f(t)|^2\mathrm{d}t = \frac{1}{2\pi}\sum_{u\in\Gamma_u}\int \left|\langle f, w_{u,\omega}\rangle\right|^2\mathrm{d}\omega$$

as soon as

$$\forall t, \quad \sum_{u\in\Gamma_u}|w(t-u)|^2 = 1.$$

Discretization of ω: If w is compactly supported in $[-\Delta,\Delta]$ then $f(t)w_{u,\omega}(t)$ is compactly supported in $[u-\Delta, u+\Delta]$. One can thus use a Fourier series decomposition to show that

$$f(t)\overline{w(t-u)} = \frac{1}{2\Delta}\sum_{k\in\mathbb{Z}} Sf(u,k2\pi/\Delta)w_{u,k2\pi/\Delta}(t)$$

while

$$\int |f(t)\overline{w(t-u)}|^2 = \frac{1}{4\Delta^2}\sum_{k\in\mathbb{Z}}|Sf(u,k2\pi/\Delta)|^2.$$

This amounts to a discretization of ω of step $2\pi/\Delta$.

Proof: We start by rewriting Sf as a convolution

$$\begin{aligned} Sf(u,\omega) &= \int f(t)\overline{w(t-u)}e^{-i\omega t}\mathrm{d}t \\ &= e^{-i\omega u}\int f(t)\overline{w(-(u-t)}e^{i\omega(u-t)}\mathrm{d}t \\ &= e^{-i\omega u} f \star (\overline{w}(-\cdot)\times e_{i\omega})(u). \end{aligned}$$

As $f \in L^2$ and $w \in L^1$, for any ω, $u \mapsto f \star (\overline{w}(-\cdot) \times e_{i\omega}) \in L^2$ and thus $u \mapsto Sf(u,\omega) \in L^2$. Let $Sf_\omega(u) = Sf(u,\omega) = e^{-i\omega u} f \star (\overline{w}(-\cdot) \times e_{i\omega})(u)$. As

$$\begin{aligned}\widehat{\overline{w}(-\cdot) \times e_{i\omega}}(\xi) &= \int \overline{w}(-t) e^{i\omega t} e^{-i\xi t} \mathrm{d}t \\ &= \int \overline{w}(t) e^{i(\xi-\omega)t} \mathrm{d}t \\ &= \overline{\int w(t) e^{-i(\xi-\omega)t} \mathrm{d}t} \\ &= \overline{\widehat{w}(\xi-\omega)},\end{aligned}$$

one derives

$$\mathcal{F}Sf_\omega = \mathcal{F}f(\cdot + \omega) \times \overline{\widehat{w}}.$$

By Parseval, one has thus for any ω,

$$\begin{aligned}\int |Sf_\omega(u)|^2 \mathrm{d}u &= \frac{1}{2\pi} \int |\mathcal{F}f(\xi+\omega)|^2 \left|\overline{\widehat{w}(\xi)}\right|^2 \mathrm{d}\xi \\ &= \frac{1}{2\pi} \int |\mathcal{F}f(\xi)|^2 \left|\overline{\widehat{w}(\xi-\omega)}\right|^2 \mathrm{d}\xi\end{aligned}$$

and thus

$$\int\int |Sf(u,\omega)|^2 \mathrm{d}u\mathrm{d}\omega = \frac{1}{2\pi} \int\int |\mathcal{F}f(\xi)|^2 \left|\overline{\widehat{w}(\xi-\omega)}\right|^2 \mathrm{d}\xi\mathrm{d}\omega$$

as everything is positive one can apply Fubini

$$\begin{aligned}\int\int |Sf(u,\omega)|^2 \mathrm{d}u\mathrm{d}\omega &= \frac{1}{2\pi} \int\int |\mathcal{F}f(\xi)|^2 \left|\overline{\widehat{w}(\xi-\omega)}\right|^2 \mathrm{d}\omega\mathrm{d}\xi \\ &= \frac{1}{2\pi} \int |\mathcal{F}f(\xi)|^2 \int \left|\overline{\widehat{w}(\xi-\omega)}\right|^2 \mathrm{d}\omega\mathrm{d}\xi\end{aligned}$$

and using again Parseval

$$= 2\pi \|f\|_2^2 \|w\|_2^2$$

which yields the energy conservation result.

Along the same lines,

$$\int Sf_\omega(u) w(t-u) e^{i\omega t} \mathrm{d}u = \frac{1}{2\pi} \int \widehat{Sf_\omega}(\xi) \overline{\widehat{w(t-\cdot)}(\xi)} \mathrm{d}\xi e^{i\omega t}.$$

Now as

$$\begin{aligned}\widehat{\overline{w(t-\cdot)}}(\xi) &= \int \overline{w(t-u)} e^{-i\xi u} \mathrm{d}u \\ &= \int \overline{w(-u)} e^{-i\xi(u+t)} \mathrm{d}u \\ &= \int \overline{w(u)} e^{i\xi(u-t)} \mathrm{d}u \\ &= e^{-i\xi t} \overline{\widehat{w}(\xi)}\end{aligned}$$

we obtain

$$\begin{aligned}\int Sf_\omega(u)w(t-u)e^{i\omega t}\mathrm{d}u &= \frac{1}{2\pi}\int \mathcal{F}(\xi+\omega)\overline{\widehat{w}(\xi)}e^{i\xi t}\widehat{w}(\xi)\mathrm{d}\xi e^{i\omega t}\\ &= \frac{1}{2\pi}\int \mathcal{F}(\xi+\omega)|\widehat{w}(\xi)|^2 e^{i(\xi+\omega)t}\mathrm{d}\xi\\ &= \frac{1}{2\pi}\int \mathcal{F}(\xi)|\widehat{w}(\xi-\omega)|^2 e^{i\xi t}\mathrm{d}\xi\end{aligned}$$

Thus

$$\int_{-A}^{A}\int Sf(u,\omega)w(t-u)e^{i\omega t}\mathrm{d}u\mathrm{d}\omega = \int_{-A}^{A}\frac{1}{2\pi}\int \mathcal{F}f(\xi)|\widehat{w}(\xi-\omega)|^2 e^{i\xi t}\mathrm{d}\xi\mathrm{d}\omega$$

Now

$$\begin{aligned}\int_{-A}^{A}\frac{1}{2\pi}\int \left|\mathcal{F}(\xi)|\widehat{w}(\xi-\omega)|^2 e^{i\xi t}\right|\mathrm{d}\xi\mathrm{d}\omega &= \frac{1}{2\pi}\int_{-A}^{A}\int |\mathcal{F}(\xi)|\,|\widehat{w}(\xi-\omega)|^2\mathrm{d}\xi\mathrm{d}\omega\\ &\leq \frac{1}{2\pi}\int_{-A}^{A}\left(\int |\mathcal{F}(\xi)\widehat{w}(\xi-\omega)|^2\,\mathrm{d}\xi\right)^{1/2}\left(\int |\widehat{w}(\xi-\omega)|^2\,\mathrm{d}\xi\right)^{1/2}\mathrm{d}\omega\end{aligned}$$

using $\|w\|_2 = 1$

$$\begin{aligned}&\leq \frac{1}{\sqrt{2\pi}}\int_{-A}^{A}\left(\int |\mathcal{F}(\xi)\widehat{w}(\xi-\omega)|^2\,\mathrm{d}\xi\right)^{1/2}\mathrm{d}\omega\\ &\leq \frac{\sqrt{2A}}{\sqrt{2\pi}}\left(\int_{-A}^{A}\int |\mathcal{F}(\xi)\widehat{w}(\xi-\omega)|^2\,\mathrm{d}\xi\mathrm{d}\omega\right)^{1/2}\\ &\leq \frac{\sqrt{2A}}{\sqrt{2\pi}}\left(\int\int |\mathcal{F}(\xi)\widehat{w}(\xi-\omega)|^2\,\mathrm{d}\xi\mathrm{d}\omega\right)^{1/2}\\ &\leq \sqrt{2A}\sqrt{2\pi}\|f\|_2^2 < +\infty\end{aligned}$$

So we can apply Fubini and obtain:

$$\int_{-A}^{A}\int Sf(u,\omega)w(t-u)e^{i\omega t}\mathrm{d}u\mathrm{d}\omega = \frac{1}{2\pi}\int \mathcal{F}f(\xi)\int_{-A}^{A}|\widehat{w}(\xi-\omega)|^2\mathrm{d}\omega e^{i\xi t}\mathrm{d}\xi$$

Let $g_A(\xi) = \int_{-A}^{A}|\widehat{w}(\xi-\omega)|^2\mathrm{d}\omega$, by construction $g_A \leq 1$ and $\lim_{A\to\infty} g_A(\xi) = 1$. Thus $\mathcal{F}f(\xi)\int_{-A}^{A}|\widehat{w}(\xi-\omega)|^2\mathrm{d}\omega = \mathcal{F}f(\xi)\times g_A(\xi)$ is in L^2 for any A. Furthermore by dominated convergence $\mathcal{F}f\times g_A$ is a Cauchy sequence which converges in L^2 toward $\mathcal{F}f$. Now

$$\int_{-A}^{A}\int Sf(u,\omega)w(t-u)e^{i\omega t}\mathrm{d}u\mathrm{d}\omega = \mathcal{F}^{-1}(\mathcal{F}f\times g_A)$$

and thus converges in L^2 toward $\mathcal{F}^{-1}\mathcal{F}f = f$.

Discretization: Given two discretization steps Δ_t and Δ_w, one can consider the countable family of atoms

$$(w_{k\Delta_t, l\Delta_w})_{(k,l)\in\mathbb{Z}^2}$$

and define the discretized windowed Fourier transform of $f \in L^2$ by

$$(\langle f, w_{k\Delta_t, l\Delta_w}\rangle)_{(k,l)\in\mathbb{Z}^2}.$$

A natural question is whether similar energy conservation property and reconstruction formula hold.

It turns out that a necessary (but not sufficient) condition is $\Delta_t \times \Delta_w \geq 2\pi$, that is a sufficiently dense sampling. We refer to Daubechies book for more details...

Finite Setting In this case, atoms are defined by

$$w_{m,l}[n] = w[n-m \mod N]e^{i2\pi\frac{ln}{N}}$$

where w is supported in $\{0,\dots,N-1\}$ such that $\sum_{n=0}^{N}|w[n]|^2 = 1$. Their Fourier transform are thus

$$\widehat{w}_{m,l}[k] = \widehat{w}[k-l]e^{i2\pi\frac{m(k-l)}{N}}$$

The discrete Windowed Fourier Transform is then defined by

$$Sf[m,l] = \sum_{n=0}^{n-1} f[n]\overline{w[n-m]}e^{-i2\pi\frac{ln}{N}} = \langle f, w_{m,l}\rangle = \frac{1}{N}\langle\widehat{f}, \widehat{w_{m,l}}\rangle = \widehat{fw[\cdot - m]}[l]$$

Note that using the FFT algorithm to compute $\widehat{fw[\cdot - m]}$ at each position n yields an algorithm of complexity $O(N^2\log N)$.

Using the same proof technique than the one used in the lapped transform interpretation, one obtains

$$f = \frac{1}{N}\sum_{m=0}^{N-1}\sum_{l=0}^{N-1} Sf[m,l]w_{m,l}$$
$$\sum_{n=0}^{N-1}|f[n]|^2 = \frac{1}{N}\sum_{m=0}^{N-1}|\sum_{l=0}^{N-1} Sf[m,l]|^2$$

which remains true if one subsamples the position on a grid of step Δ as soon as one imposes

$$\sum_{k=0}^{k<\frac{N}{\Delta}} |w[k\Delta + n \mod N]|^2 = 1, \forall n$$

3.3 Instantaneous Frequency

In a musical score, one can *see* several *frequencies* that vary along time. To formalize this idea, one need first to define the notion on *Instantaneous frequency.*

3.3.1 Definition

The cos **cas:** Let f be a modulated cosine

$$f(t) = a\cos(\omega_0 t + \phi_0) = a\cos\phi(t)$$

its frequency ω_0 measures the *speed* of rotation of the angle in the cos, which in nothing but the derivative of the angle ϕ.

A first (attempt of) generalization: Let f be a real valued signal

$$f(t) = a(t)\cos\phi(t)$$

with $a(t) \geq 0$ and $\phi'(t) \geq 0$, one could define its instantaneous frequency $\omega(t)$ by

$$\omega(t) = \phi'(t).$$

This definition is not satisfying because the decomposition $f(t) = a(t)\cos\phi(t)$ is not unique and thus there is no unique way of defining this instantaneous frequency.

An analytic fix: Let f be a real valued signal, as its spectrum has an hermitian symmetry, it is entirely characterized by its Fourier transform restricted to the positive axis. We define the *analytic part* f_a of f has the function whose Fourier transform is defined by

$$\widehat{f_a}(\omega) = \begin{cases} 2f(\omega) & \text{if } \omega \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

By construction,

$$\widehat{f}(\omega) = \frac{1}{2}\left(\widehat{f_a}(\omega) + \overline{\widehat{f_a}(-\omega)}\right)$$

and thus

$$f = \frac{1}{2}\left(f_a + \overline{f_a}\right) = \Re(f_a).$$

The signal f_a is called analytic because it has an analytic extension in the upper half plane. Now, as

$$f_a(t) = |f_a(t)|e^{i\,\mathrm{Arg}\, f_a(t)}$$

we obtain

$$f(t) = |f_a(t)|\cos \mathrm{Arg}\, f_a(t).$$

One says that $a(t) = |f_a(t)|$ is the *analytic* amplitude of $f(t)$ while $\omega(t) = (\mathrm{Arg}\, f_a)'(t)$ is its *analytic* instantaneous frequency. Both are defined now in a unique way.

Example: If $f(t) = a(t)\,\cos(\omega_0 t + \phi_0)$ then

$$\widehat{f}(\omega) = \frac{1}{2}\left(e^{i\phi_0}\,\widehat{a}(\omega - \omega_0) + e^{-i\phi_0}\,\widehat{a}(\omega + \omega_0)\right).$$

Assume that $\widehat{a}$ is supported in $[-\omega_0, \omega_0]$, i.e.that a varies slowly with respect to the period $2\pi/\omega_0$ of the cos, then

$$\widehat{f_a}(\omega) = e^{i\phi_0}\,\widehat{a}(\omega - \omega_0)$$

and thus $f_a(t) = a(t)\,e^{i(\omega_0 t + \phi_0)}$.

Frequency Modulation: In order to transmit a band limited signal $m(t)$, one transmits

$$f(t) = a\cos(\omega_0 t + k\int_0^t m(u)\mathrm{d}u).$$

If ω_0/k is large with respect to the Nyquist frequency of m, then the instantaneous frequency of f is very close from being equal to $\omega_0 + km(t)$. The signal m can thus be recovered from f. This methodology is more noise resistant and power efficient than AM.

Limit of the instantaneous frequency: If f is the sum of two cosines:

$$f(t) = a\cos\omega_1 t + a\cos\omega_2 t$$

then

$$\begin{aligned} f_a(t) &= ae^{i\omega_1 t} + ae^{i\omega_2 t} \\ &= a\cos\left(\frac{\omega_1-\omega_2}{2}t\right)e^{i\frac{\omega_1+\omega_2}{2}t}. \end{aligned}$$

The definition is thus interesting only if there is one single frequency (or they are well separated...)

3.3.2 Instantaneous frequency and Windowed Fourier transform

It turns out that the instantaneous frequency can be read in the Windowed Fourier Transform or in its spectogram. Let w be a window in $L^1 \cap L^2$, with $\|w\|_2 = 1$, centered in time and frequency and such that $\widehat{w}(\omega) \ll 1$ as soon as $|\omega| \geq \Delta$. We define its scaled version by

$$w_s(t) = \frac{1}{\sqrt{s}}w(t/s)$$

and let the Windowed Fourier Transform depends on s:

$$Sf(s,u,\omega) \ = \int_{-\infty}^{+\infty} a(t)\cos\phi(t)\overline{w_s}(t-u)\,e^{-i\omega t}\mathrm{d}t = \langle f, w_{s,u,\omega}\rangle.$$

Theorem 3.3: Let $f(t) = a(t)\cos\phi(t)$. If the variations of $a(t)$ and $\phi'(t)$ are small on the support $[u-s, u+s]$ of $w_s(\cdot - u)$

$$\int_{-\infty}^{+\infty} \left|a(t+u)e^{i\phi(t+u)} - a(u)e^{i\left(\phi(u)+\phi'(u)t\right)}\right| |\overline{w_s}(t)|\mathrm{d}t \ll a(u)\sqrt{s}\|w\|_1$$

and if $\phi'(u) \geq s^{-1}\,\Delta$ then for all $\omega \geq 0$,

$$Sf(s,u,\omega) \approx \frac{\sqrt{s}}{2}a(u)e^{i(\phi(u)-\omega u)}\widehat{w}\Big(s[\omega-\phi'(u)]\Big).$$

Proof: As

$$\begin{aligned} Sf(s,u,\omega) &= \langle f, w_{s,u,\omega}\rangle \\ &= \int_{-\infty}^{+\infty} a(t)\cos\phi(t)\overline{w_s}(t-u)\,e^{-i\omega t}\mathrm{d}t \\ &= \frac{1}{2}\int_{-\infty}^{+\infty} a(t)\,\left(e^{i\phi(t)} + e^{-i\phi(t)}\right)\overline{w_s}(t-u)e^{-i\omega t}\mathrm{d}t \\ &= \frac{1}{2}\left(I(\phi) + I(-\phi)\right). \end{aligned}$$

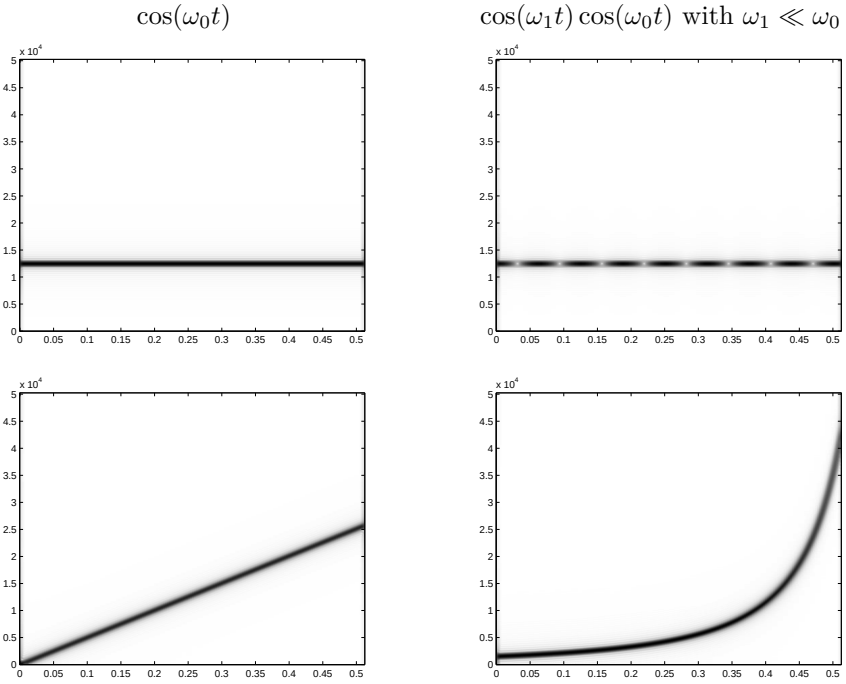


Figure 3.7: STFT

Now

$$\begin{aligned}
I(\phi) &= \int_{-\infty}^{+\infty} a(t) e^{i\phi(t)}\, \overline{w_s}(t-u)\, e^{-i\omega t} \mathrm{d}t \\
&= \int_{-\infty}^{+\infty} a(t+u) e^{i\phi(t+u)}\, \overline{w_s}(t)\, e^{-i\omega(t+u)} \mathrm{d}t \\
&= \int_{-\infty}^{+\infty} a(u) e^{i\left(\phi(u)+\phi'(u)t\right)}\, \overline{w_s}(t)\, e^{-i\omega(t+u)} \mathrm{d}t \\
&\quad + \int_{-\infty}^{+\infty} \left(a(t+u) e^{i\phi(t+u)} - a(u) e^{i\left(\phi(u)+\phi'(u)t\right)} \right) \overline{w_s}(t)\, e^{-i\omega(t+u)} \mathrm{d}u \\
&\approx \int_{-\infty}^{+\infty} a(u) \overline{w_s}(t) e^{i(\phi(u)-\omega u)} e^{-i(\omega-\phi'(u))t} \mathrm{d}t \\
&\approx a(u) e^{i(\phi(u)-\omega u)} \widehat{\overline{w_s}}(\omega - \phi'(u)) \\
&\approx a(u) e^{i(\phi(u)-\omega u)} \sqrt{s} \overline{\widehat{w}(s(\phi'(u)-\omega))}
\end{aligned}$$

Now using the same computation

$$I(-\phi) \approx a(u) e^{i(\phi(u)-\omega u)} \sqrt{s} \overline{\widehat{g}(s(-\phi'(u)-\omega))}$$

As $\omega \geq 0, s|\omega + \phi'(u)| > \Delta$ and thus $|\overline{\widehat{g}(s(-\phi'(u)-\omega))}| \ll 1$ and $I(-\phi) \ll |a(u)|$.

Ridges: Under the local assumptions of the previous theorem, the spectrogram

$$Pf(s,u,\omega) = |Sf(s,u,\omega)|^2 = |\langle f, w_{s,u,\omega}\rangle|^2$$

of $f(t) = a(t)\cos\phi(t)$ is approximately

$$Pf(s,u,\omega) \approx \frac{s}{4}|a(u)|^2 \left|\widehat{w}\left(s[\omega - \phi'(u)]\right)\right|^2.$$

Under the mild assumption that $|\widehat{w}|$ is maximal at 0, the spectrogram is thus maximum at $\omega(u) = \phi'(u)$, the instantaneous frequency. The points $(u, \omega(u))$ are called ridges.

At those points,

$$Sf(u,\xi(u)) \approx \frac{\sqrt{s}}{2} a(u) e^{i(\phi(u)-\xi u)}\widehat{w}(0)$$

so that

$$a(u) \approx \frac{2}{\sqrt{s}\,|\widehat{w}(0])|}|Sf(s,u,\omega(u)|.$$

Furthermore,

$$\frac{\partial}{\mathrm{d}u}(\phi(u) - \xi u)(u,\xi(u)) = \phi'(u) - \xi(u) \approx 0$$

and thus the phase is approximately constant along the ridges.

3.3.3 Multiple frequencies:

If a function is a finite sum of terms $a_k(t)\cos\phi_k(t)$

$$f = \sum_{k=1}^{K} a_k(t)\cos\phi_k(t)$$

then the Windowed Fourier Transform will allow to separate those terms, provided their instantaneous frequencies are sufficiently separated. More precisely, assume the assumption of the previous theorem holds for all couples (a_k, ϕ_k) then

$$Sf(s,u,\omega) \approx \frac{\sqrt{s}}{2}\sum_{k=1}^{K} a_k(u) e^{i(\phi_k(u)-\omega u)}\widehat{w}\Big(s[\omega - \phi_k'(u)]\Big).$$

Separated ridges: As soon as $\min_{k'\neq k}|\phi_k'(u) - \phi_{k'}'(u)| \geq \frac{\Delta}{s}$ then in the neighborhood of $\phi_k'(u)$,

$$Sf(s,u,\omega) \approx \frac{\sqrt{s}}{2} a_k(u) e^{i(\phi_k(u)-\omega u)}\widehat{w}\Big(s[\omega - \phi_k'(u)]\Big)$$

and thus a ridge corresponding to $a_k\cos\phi_k(t)$ exists.

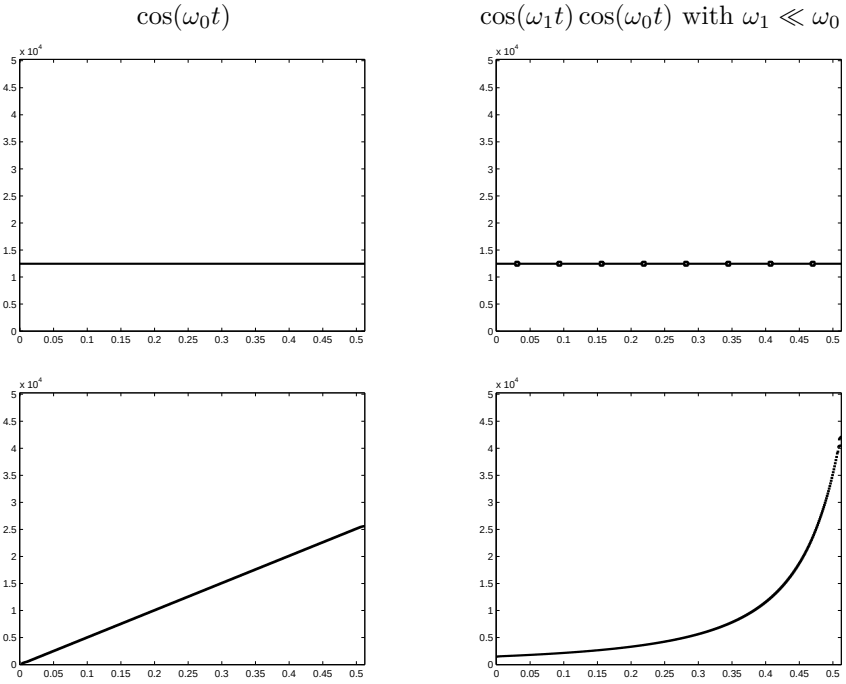


Figure 3.8: Ridges

Interferences: When $\min_{k' \neq k} |\phi'_k(u) - \phi'_{k'}(u)| < \frac{\Delta}{s}$, the previous analysis cannot be used. The two ridges are not separated anymore, and interference patterns may be observed.

Window scale choice: The scale s of the window is a crucial parameter: on the one hand it has to be small so that the local approximation of $a(t)\cos\phi(t)$ by $a(u)\cos(\phi(u) + \phi'(u)(t-u))$ is valid on the neighborhood of u of size $O(s)$, on the other hand it has to be large so that the frequencies could be resolved, as the separation limit is given by Δ/s.

Additive synthesis: Assume we analyze a sound f and that one detects K ridges, the sound f is then naturally modeled by

$$\sum_{k=1}^{K} a_k(t)\cos\phi_k(t).$$

This representation is particularly well suited to scale the duration of a sound without modifying its tone or the other way around. Indeed, the ear is sensitive to the amplitude and the instantaneous frequency, which are quantities that should be preserved.

A sound time scaled by a factor α can thus be defined by

$$\sum_{k=1}^{K} a_k(t/\alpha)\cos\left(\alpha\phi_k(t/\alpha)\right)$$

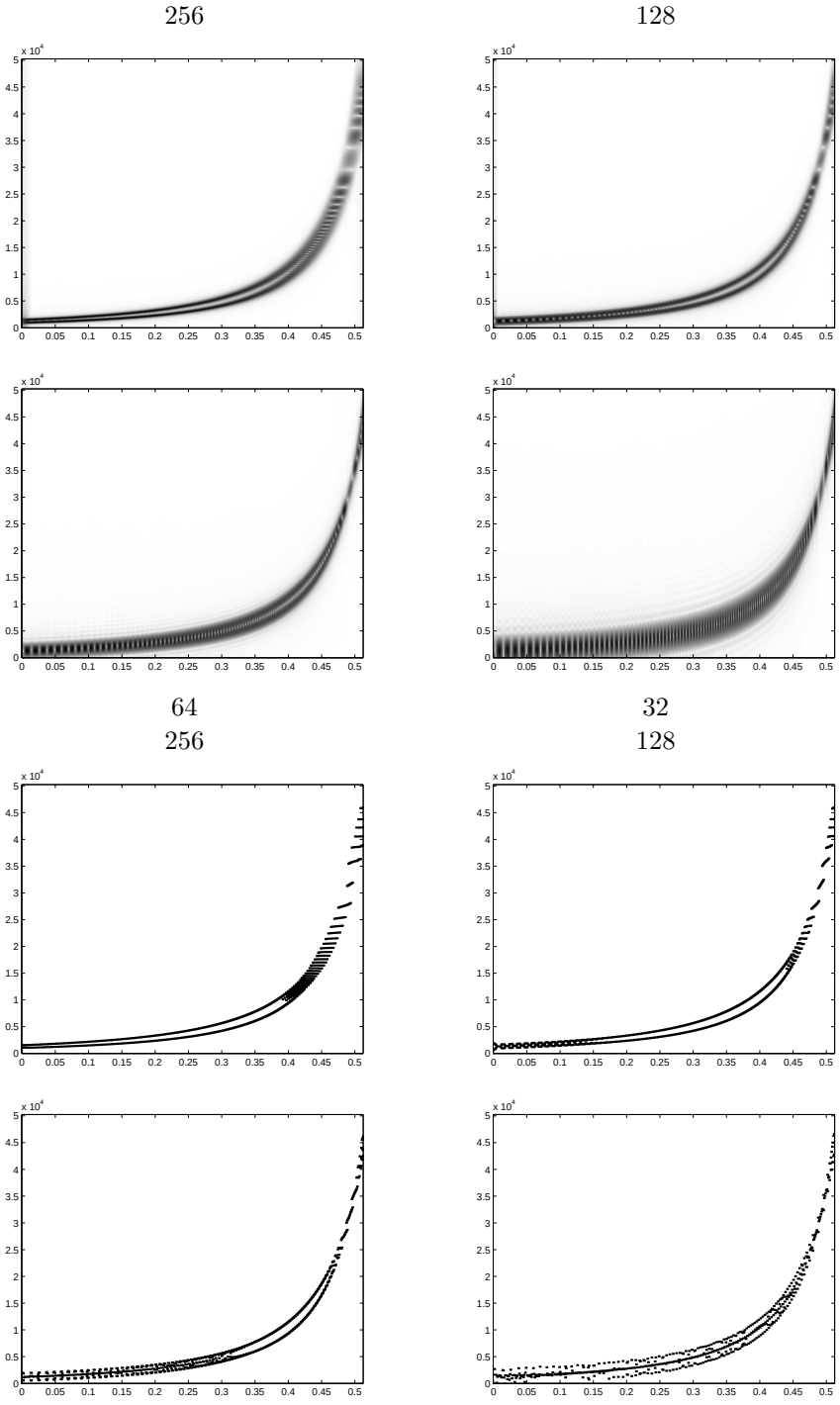


Figure 3.9: STFT and ridges for various window sizes (in samples)

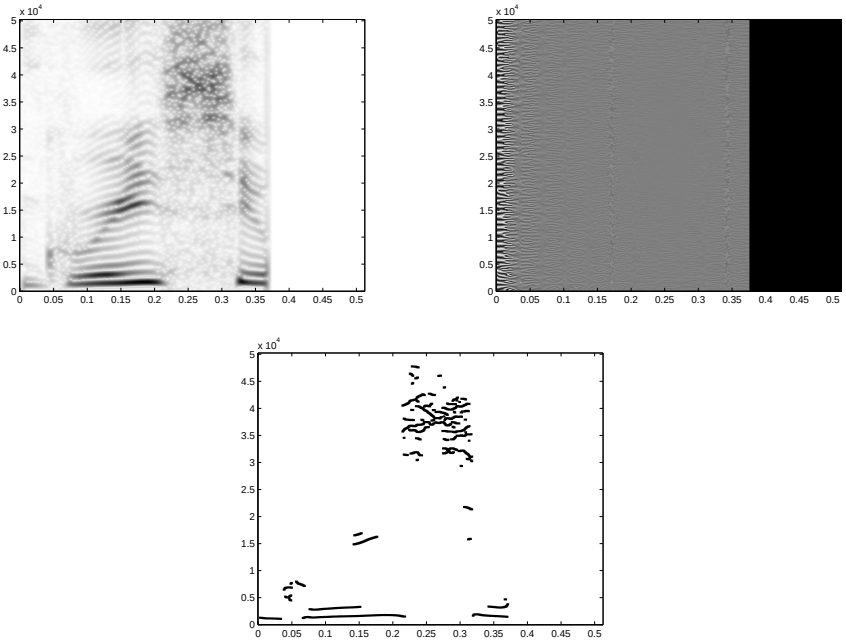


Figure 3.10: Ridge detection for Greasy sound

while a sound frequency scaled by a factor α can be defined by

$$\sum_{k=1}^{K} a_k(t) \cos\left(\alpha \phi_k(t)\right).$$

Note that more advanced modeling of the amplitude and the phase is possible.

Chapter 4

Discrete Stationary Random Process

In the previous figure, a *chaotic* part seems to exist. In order to model such a behavior, we will go to random modeling.

We consider here discrete sequences in a random process setting. This is the natural setting when there is uncertainty in the measures or when one wants to model the variability of a signal class.

We will restrict ourself to some simple processes: the wide-sense stationary processes.

4.1 Wide sense stationary processes

4.1.1 Definition

A discrete random process X is a sequence of random variables $X[n]$ with $n \in \mathbb{Z}$. We restrict ourself to real or complex random variables. Such a process is entirely characterized by the joint laws of finite random vectors extracted from X: it suffices to know $\forall k > 0$, $\forall (n_1, \ldots, n_k) \in \mathbb{Z}^k$, $\forall (A_1, \ldots, A_k) \in \mathcal{B}^k$,

$$P(X[n_1] \in A_1, \ldots X[n_k] \in A_k),$$

where $\mathcal{B}$ is the set of Borel sets.

Stationarity: A random process is said to be strictly stationary if, $\forall (n_1, \ldots, n_k) \in \mathbb{Z}^k$ and any $\delta \in \mathbb{Z}$ the law of $(X[n_1 + \delta], \ldots, X[n_k + \delta])$ is the law of $(X[n_1], \ldots, X[n_k])$

We restrict ourself to the important class of random processes having finite moments of order 2,

$$\forall n, \mathbb{E}\left(|X[n]|^2\right) < +\infty.$$

Those processes are such that $\mathbb{E}(X[n])$ exists and is finite as

$$|\mathbb{E}(X[n])| \leq \mathbb{E}(|X[n]|^2)^{1/2}\mathbb{E}(1)^{1/2} = \mathbb{E}(|X[n]|^2)^{1/2}.$$

Definition: Let X be a process having finite moment of order 2, the discrete function R_X defined by

$$\forall (n,m) \in \mathbb{Z}^2, \quad R_X[n,m] = \mathrm{Cov}(X[n], X[m]) = \mathbb{E}\left((X[n] - \mathbb{E}X[n]) \, \overline{(X[m] - \mathbb{E}X[m])} \right)$$

is called the autocovariance function of X.

The finite moment of order 2 assumption ensures that this function is well defined.

Definition: A discrete wide-sense stationary process (WSSP) is a discrete process X such that $\forall n \in \mathbb{Z}$ $X[n]$ has a finite moment of order 2 and satisfies

- $\forall n \in \mathbb{Z},\ \mathbb{E}X[n] = \mathbb{E}X[0]$,
- $\forall (n,m) \in \mathbb{Z},\ R_X[n,m] = R_X[n-m, 0]$.

If X is a WSSP, with a slight abuse of notation, the autocovariance R_X is seen as discrete function of a single variable: $R_X[n] = R_X[n, 0]$.

Gaussian case: If a discrete process is Gaussian then its wide-sense stationarity implies its strict-sense stationarity!

Proposition 4.1 (WSSP):

i) Let X_0 be a random process having a finite moment of order 2, the random process X defined $\forall n \in \mathbb{Z} X[n] = X_0$ is a WSSP.

ii) Let $(Y[n])$ be an i.i.d. random variables sequence having finite moment of order 2, the process Y is a WSSP.

iii) Let $(Z[n])$ be a sequence of decorrelated random variables having the same mean and the same variance, the process Z is a WSSP.

Proof: i) X is a process such that $\forall n \in \mathbb{Z}, X[n]$ has a finite moment of order 2, $EX[n] = EX_0 = EX[0]$ and $\forall (n,m) \in \mathbb{Z}^2$

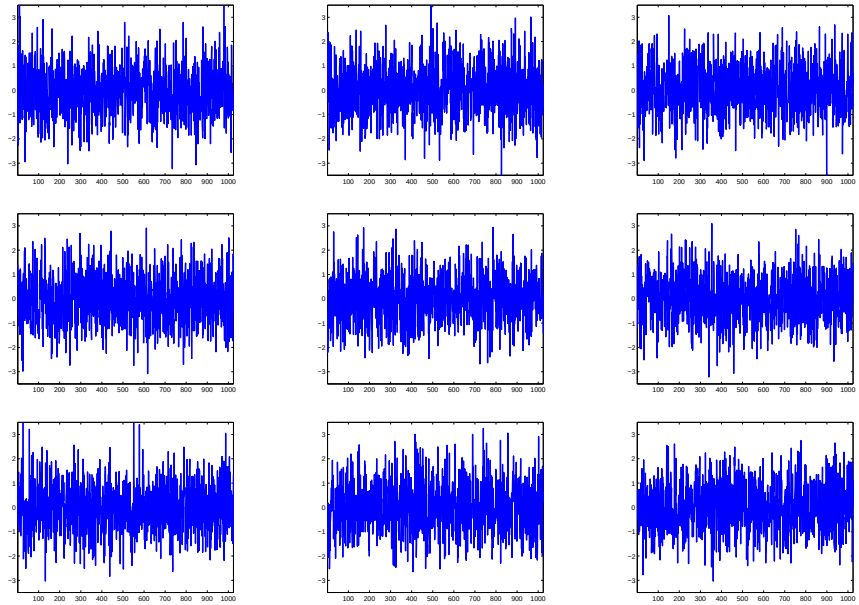
$$R_X[n,m] = \mathrm{Cov}(X[n], X[m]) = \mathrm{Cov}(X_0, X_0) = R_X[n-m, 0].$$

ii) Y is a process such that $\forall n \in \mathbb{Z}, Y[n]$ has a finite moment of order 2, and such that, using the i.i.d. property, $\forall n \in \mathbb{Z}$; $EY[n] = EY[0]$ and $\forall (n,m) \in \mathbb{Z}^2$

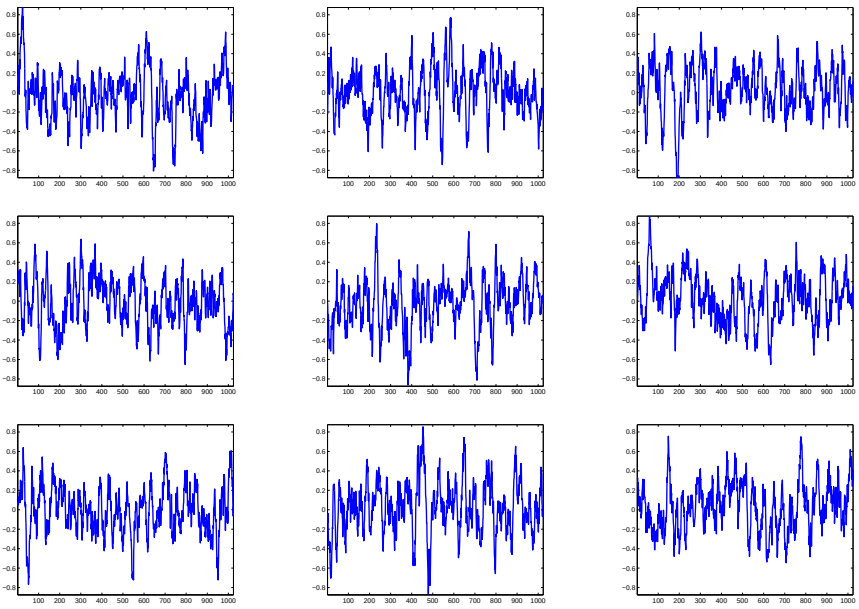
$$R_Y[n,m] = \mathrm{Cov}(Y[n], Y[m]) = \mathrm{Cov}(Y[0], Y[0])\delta[n-m] = R_Y[n-m].$$

iii) The result is obtain with the same computations than for case ii)

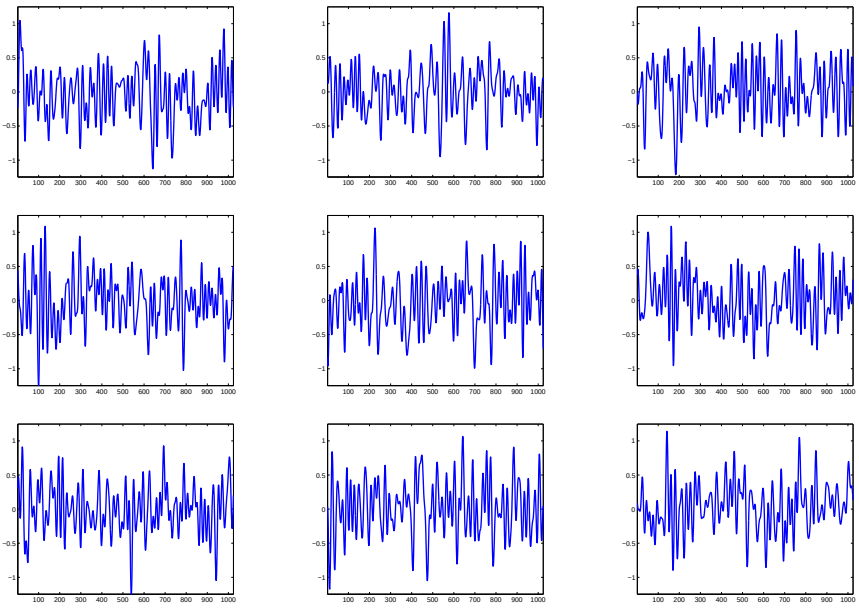
White noise: A zero mean WSSP with autocovariance function $\sigma^2\delta$ is called a white noise.



W



S



L

4.1.2 Autocovariance function properties

Proposition 4.2: The autocovariance function R_X of a WSSL X satisfies

- $R_X[0] \in \mathbb{R}^+$,
- $\forall n \in \mathbb{Z},\ R_X[-n] = \overline{R_X[n]}$
- $\forall n \in \mathbb{Z},\ |R_X[n]| \leq R_X[0]$

Proof: The proof of the first item is immediate: $R_X[0] = \mathrm{Cov}(X[0], X[0]) \geq 0$ as a variance is non negative.

The proof of the second item is obtained from the definition of the autocovariance and the stationarity of the process

$$R_X[-n] = \mathrm{Cov}(X[-n], X[0]) = \overline{\mathrm{Cov}(X[0], X[-n])} = \overline{\mathrm{Cov}(X[n], X[0])} = \overline{R_X[n]}$$

The last item is proved in a similar way that the Cauchy-Schwartz inequality to which it corresponds.If $R_X[n] = 0$ then the property holds. Otherwise, $\forall \lambda \in \mathbb{C}$,

$$\begin{aligned}\mathrm{Cov}(X[0] + \lambda X[n], X[0] + \lambda X[n]) &= \mathrm{Cov}(X[0], X[0) + \overline{\lambda}\,\mathrm{Cov}(X[0], X[n]) \\ &\quad + \lambda\,\mathrm{Cov}(X[n], X[0]) + |\lambda|^2\,\mathrm{Cov}(X[n], X[n]) \\ &= (1 + |\lambda|^2)R_X[0] + (\lambda R_X[n] + \overline{\lambda R_X[n]} \geq 0.\end{aligned}$$

Applying this result to $\lambda = \mu \frac{\overline{R_X[n]}}{|R_X[n]|}$ with μ real yields $\forall \mu \in \mathbb{R}, (1 + \mu^2)R_X[0] + 2\mu|R_X[n]| \geq 0$. The discriminant of this real polynomial of degree 2 is non positive, that is $|R_X[n]|^2 \leq R_X[0]^2$.

Theorem 4.1: The autocovariance function R_X of a WSSP is a discrete function of positive type, i.e. $\forall k \in \mathbb{N}^*, \forall (n_1 \ldots, n_k) \in \mathbb{Z}, \left[R_X[n_i - n_j]\right]_{1 \leq i,j \leq k}$ in a positive semi definite hermitian matrix:

$$\forall k \in \mathbb{N}^*, \forall (n_1 \ldots, n_k) \in \mathbb{Z}, \forall (\xi_1, \ldots, \xi_k) \in \mathbb{C}^k, \sum_{1 \leq i,j \leq k} \xi_i R_X[n_i - n_j]\overline{\xi_j} \in \mathbb{R}^+.$$

Proof: It suffices to notice that

$$\begin{aligned}\sum_{1 \leq i,j \leq k} \xi_i R_X[n_i - n_j]\overline{\xi_j} &= \sum_{1 \leq i,j \leq k} \xi_i\,\mathrm{Cov}(X[n_i], X[n_j]\overline{\xi_j}) \\ &= \mathrm{Cov}\left(\sum_{1 \leq i \leq k} \xi_i X[n_i], \sum_{1 \leq i \leq k} \xi_i X[n_i]\right)\end{aligned}$$

is a variance and thus non negative.

4.1.3 Ergodicity, mean and covariance estimation

Setting: Assume we observe a sample of N succesives values $X[0], \ldots, X[N-1]$ of a WSSP X. One may be interested in estimating its mean μ and its covariance function R_X from those observation.

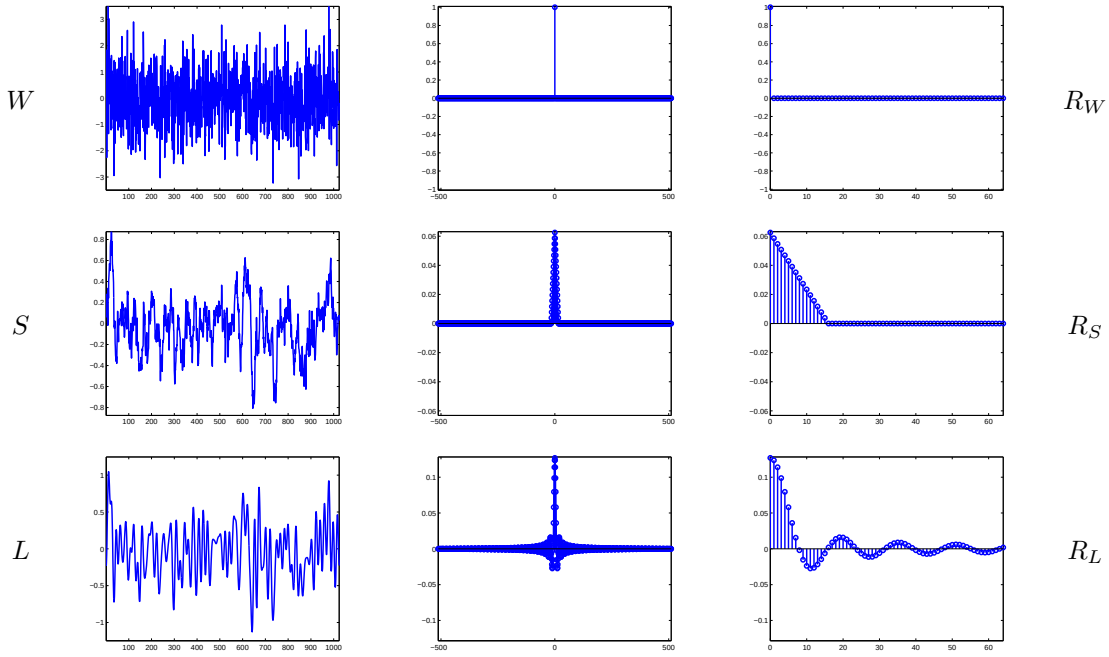


Figure 4.2: Autocovariance functions (with some zoom around 0)

Mean estimation: The most natural estimate is the empirical mean of the process

$$\widetilde{\mu} = \frac{1}{N} \sum_{k=0}^{N} X[k].$$

While unbiased ($\mathbb{E}\widetilde{\mu} = \mu$), it is not necessarily consistent when N goes to $+\infty$. For instance, if $X[k] = X[0]$ then $\widetilde{\mu} = X[0]$...

Consistency and ergodicity: If $\widetilde{\mu}$ is consistent for the quadratic loss, the process is said to be ergodic for the mean.

Proposition 4.3: the process is ergodic for the mean if only if

$$\lim_{N\to\infty} \frac{1}{N} \sum_{l=-N+1}^{N-1} \left(1 - \frac{l}{N}\right) R_X[l] = 0$$

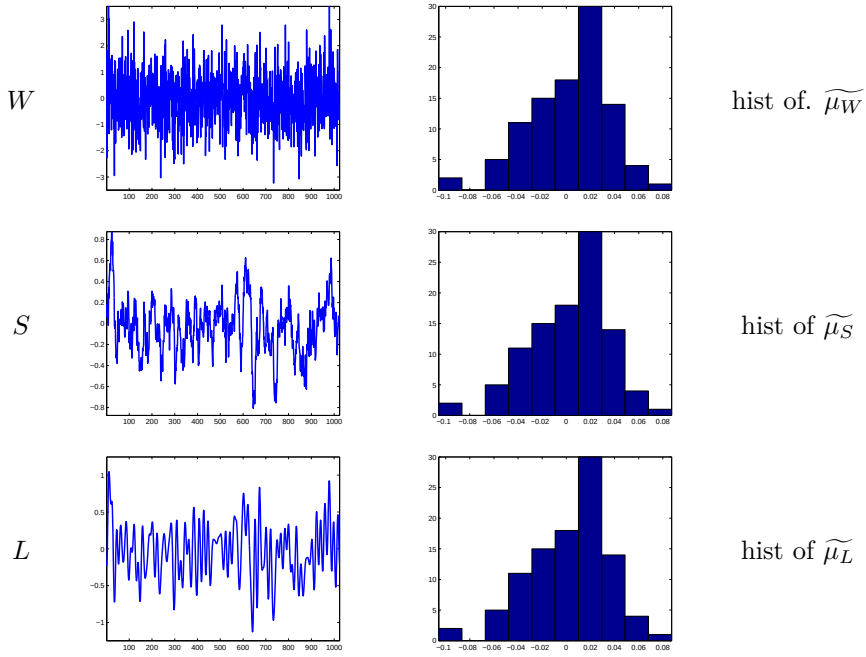


Figure 4.3: Mean estimation

Proof:

$$\begin{aligned}
\mathbb{E}\left(|\tilde{\mu}_N - \mu|^2\right) &= \frac{1}{N^2}\mathbb{E}\left(\sum_{0\leq n,m} N-1 (X[n]-\mu)\overline{(X[m]-\mu)}\right) \\
&= \frac{1}{N^2}\sum_{0\leq n,m\leq N-1} R_X[n-m] \\
&= \frac{1}{N^2}\sum_{-N+1\leq l\leq N-1} (N-|l|)R_X[l] \\
\mathbb{E}\left(|\tilde{\mu}_N - \mu|^2\right) &= \frac{1}{N}\sum_{n=-N+1}^{N-1}\left(1-\frac{|n|}{N}\right)R_X[n]
\end{aligned}$$

Corollary 4.1: The process is ergodic for the mean if $\lim_{|k|\to\infty} R_X[k] = 0$.

Classical covariance estimate: $\forall 0 \leq k \leq N-1$

$$\widetilde{R}_X[k] = \frac{1}{N}\sum_{l=0}^{N-1-k}(X[n]-\widetilde{\mu})\overline{(X[n-k]-\widetilde{\mu}}$$

Although this estimate is biased, it is a positive type function and thus very useful in practice. Proving its consistency is more challenging and require some technical assumtions on the process.

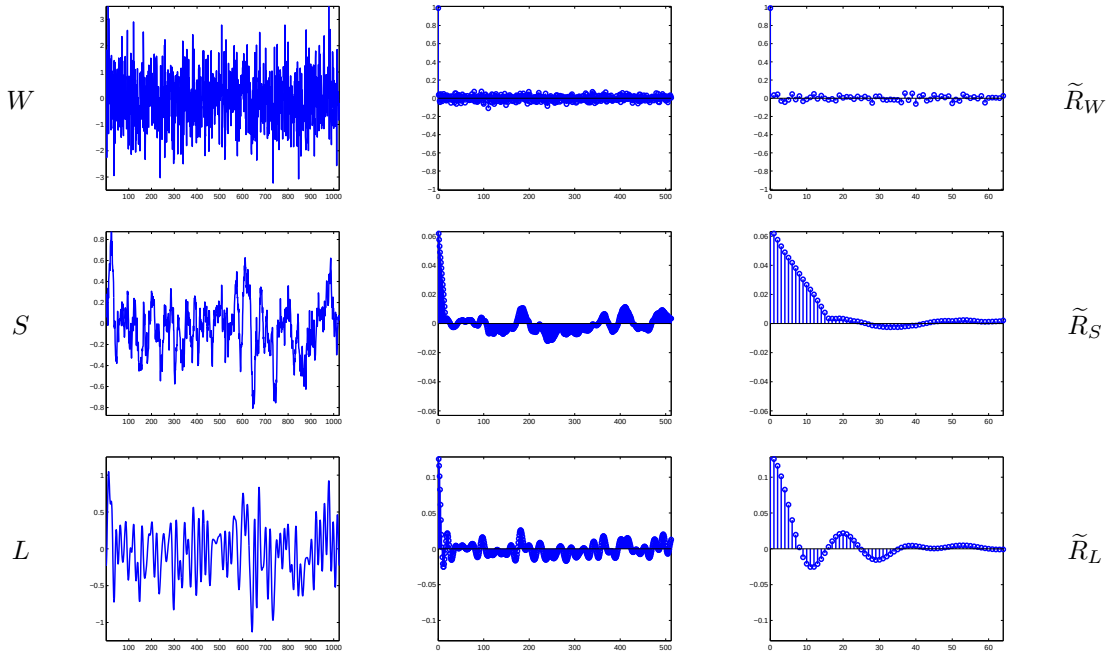


Figure 4.4: Covariance estimate

4.1.4 Autocovariance function and spectral measure

Theorem 4.2 (Herglotz): If R is a discrete function of positive type, then it exists a unique positive measure μ on $[-\pi, \pi]$ such that

$$\forall n \in \mathbb{Z}, R[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{in\omega} d\mu(\omega).$$

Remark: This is nothing but a specific Fourier transform representation.

Corollary 4.2: If $R \in \ell^1$, $\mu = \hat{R}(e^{i\omega})d\omega$ where

$$\hat{R}(e^{i\omega}) = \sum_{n \in \mathbb{Z}} R[n] e^{-in\omega}.$$

Remark: In ℓ^1, one can verify that $\widehat{h \star g} = \hat{h}\hat{g}$.

Proof: Let

$$\begin{aligned} \forall N \geq 0, \widehat{R_N}(e^{i\omega}) &= \frac{1}{N} \sum_{0 \leq m,n \leq N-1} e^{-in\omega} e^{im\omega} R[n-m] \geq 0 \text{ by assumption.} \\ &= \sum_{n=-N+1}^{N-1} (1 - |n|/N) R_X[n] e^{-in\omega} \end{aligned}$$

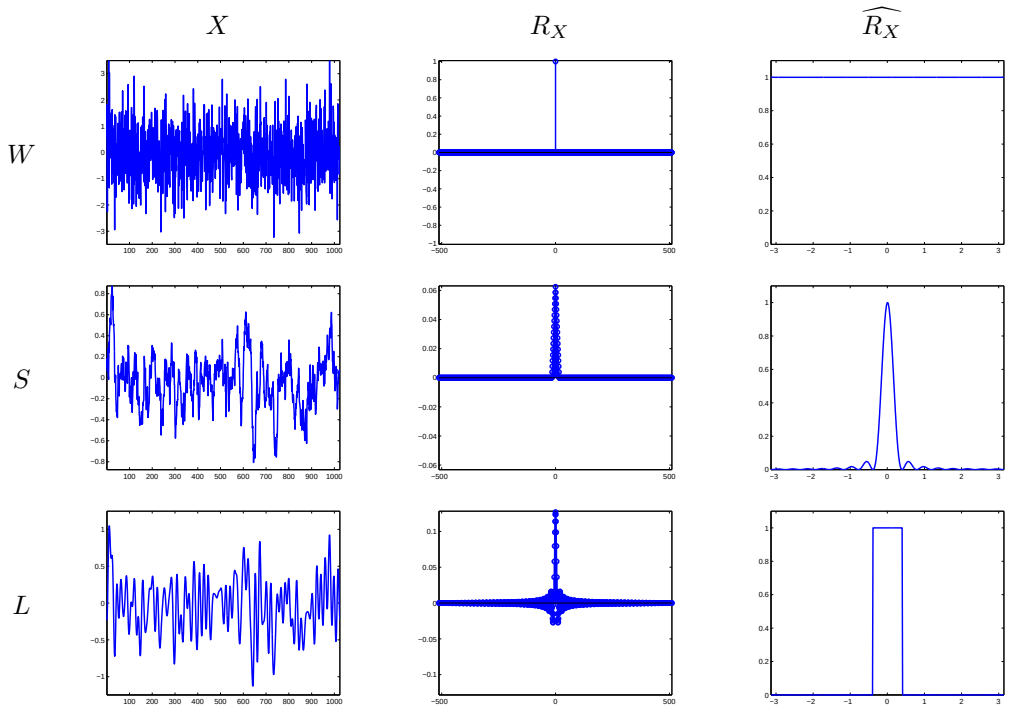


Figure 4.5: Power spectral density

The measure $d\mu_N(e^{i\omega}) = \widehat{R_N}(e^{i\omega})\mathrm{d}\omega$ is then non negative and satisfies $\forall n \in \mathbb{Z}$,

$$(1 - |n|/N)_+ R[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{in\omega} d\mu_N(\omega).$$

It suffices then to let N goes to infinity and to extract from $\mu_N([-\pi, \pi]) = 2\pi R_X[0]$ a subsequence converging tightly to a positive measure μ which satisfies the equality of the theorem.

Uniqueness is obtained by observing that the family $(e^{in\omega})_{n\in\mathbb{Z}}$ is dense into the bounded continuous functions of $[-\pi, \pi]$ and thus a measure is entirely defined by its action on this family.

Definition: Let X be a WSSP, the measure μ_X associated to the autocovariance function R_X is called the spectral measure of the process. If $R_x \in \ell^1$, the Fourier transform of R_X $\hat{R}_X(e^{i\omega})$ is called the spectral density, or the power spectral density, of X.

Remark: The use of the word *power* will be justified later.

4.1.5 Process filtering

Theorem 4.3: Let X be a WSSP and $h \in \ell^1$, the process $Y = h \star X$ defined by, $\forall n \in \mathbb{Z}$,

$$Y[n] = \sum_{k\in\mathbb{Z}} h[k]X[n-k]$$

is a WSSP of autocovariance function

$$R_Y = h \star \tilde{h} \star R_X,$$

where $\tilde{h}$ is defined $\forall n \in \mathbb{Z}, \tilde{h}[n] = \overline{h[-n]}$. Moreover, the spectral measure μ_Y of Y is given by $d\mu_Y(\omega) = |\hat{h}(e^{i\omega})|^2 d\mu_X(\omega)$.

Remark: If $\mathbb{E}(X[0]) = 0$ and $R_X \in \ell^1$ then $h \in \ell^2$ is sufficient for the well definition of $h \star X$.

Proof: For sake of simplicity, we denote for any random variable Z having a finite moment of order 2 $\|Z\|^2 = \mathbb{E}(|Z|^2)$ this moment and talk of L^2 convergence for a convergence in this norm..

Given $n \in \mathbb{Z}$, for any $k \in \mathbb{Z}$, the random variables $h[k]X[n-k]$ in the sum defining $Y[n]$ satisfy

$$\|h[k]X[n-k]\|_1 \leq |h[k]|\|X[n-k]\|_2 = |h(k)|\sqrt{|\mathbb{E}X[0]|^2 + R_X[0]}$$

because X is a WSSP. So

$$\sum_{k\in\mathbb{Z}} \|h[k]X[n-k]\|_1 \leq \sum_{k\in\mathbb{Z}} \|h[k]X[n-k]\|_2 = \|h\|_{\ell^1}\sqrt{|\mathbb{E}X[0]|^2 + R_X[0]} < +\infty.$$

This implies the almost everywhere convergence of $\sum_{k\in\mathbb{Z}} |h|[k]|X|[n-k]$ and thus the one of $\sum_{k\in\mathbb{Z}} h[k]X[n-k]$. The random variable $Y[n]$ is thus well defined.

The bound $\sum_{k\in\mathbb{Z}} \|h[k]X[n-k]\|_1$ implies that $\mathbb{E}|Y[n]| = \|Y\|_1$ is finite and that a

$$\mathbb{E}Y[n] = \sum_{k\in\mathbb{Z}} h[k]\mathbb{E}X[n-k] = \mathbb{E}X[0]\sum_{k\in\mathbb{Z}} h[k].$$

The mean of $Y[n]$ is thus independent of n.

In a similar way, the bound on $\sum_{k\in\mathbb{Z}} \|h[k]X[n-k]\|_2$ implies that $\mathbb{E}|Y[n]|^2 < +\infty$. It remains to prove the formula for the variance. For any $N \in \mathbb{N}^\star$, we define

$$Y_N[n] = \sum_{k=-N}^{N} h[k]X[n-k]$$

which converges to Y in L^2 when N goes to infinity,

$$\begin{aligned}
\mathrm{Cov}(Y_N[n], Y_N[m]) &= \mathrm{Cov}\left(\sum_{k=-N}^{N} h[k]X[n-k], \sum_{k=-N}^{N} h[k]X[m-k]\right) \\
&= \sum_{k=-N}^{N}\sum_{l=-N}^{N} h[k]\overline{h[l]}\,\mathrm{Cov}(X[n-k], X[m-l]) \\
&= \sum_{k=-N}^{N} h[k] \sum_{l=-N}^{N} \tilde{h}[l]R_X[(n-m)-k-l]
\end{aligned}$$

thus letting $h_N[k'] = h[k']\mathbf{1}_{|k'|\leq N}$

$$\mathrm{Cov}(Y_N[n], Y_N[m]) = (h_N \star \tilde{h}_N \star R_X)[n-m].$$

By L^2 continuity of the covariance operator $\mathrm{Cov}(Y_N[n], Y_N[m])$ tends to $\mathrm{Cov}(Y[n], Y[m])$. For the convolution result, it suffices to notice that

- R_X is in ℓ^∞,
- h_N tends to h in ℓ^1
- and the convolution of a ℓ^∞ discrete function by a ℓ^1 discrete function is ℓ^∞ discrete function and this application is continuous with respect to the two variables

to obtain $(h_N \star \tilde{h}_N \star R_X)[n-m] \xrightarrow[N\to+\infty]{} (h \star \tilde{h} \star R_X)[n-m]$. This concludes the proof that Y is a WSSP and that the covariance formula holds.

The spectral measure associated to Y_N can be obtained from the previous formula:

$$R_{Y_N}[n] = \sum_{k=-N}^{N} \sum_{l=-N}^{N} h[k]\overline{h[l]} R_X[n-k+l]$$

which yields by inserting the spectral measure μ_X of X

$$\begin{aligned} &= \sum_{k=-N}^{N} \sum_{l=-N}^{N} h[k]\overline{h[l]} \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i(n-k+l)\omega} d\mu_X(\omega) \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{in\omega} \sum_{k=-N}^{N} \sum_{l=-N}^{N} h[k] e^{-ik\omega} \overline{h[l] e^{-il\omega}} d\mu_X(\omega) \\ R_{Y_N}[n] &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{in\omega} |\hat{h}_N(e^{i\omega})|^2 d\mu_X(\omega). \end{aligned}$$

By L^2 continuity of the covariance, $R_{Y_N}[n]$ tends to $R_Y[0]$. The sequence h_N tends to h in ℓ^1, which implies that $\hat{h}_N$ tends uniformly to $\hat{h}$ which is a bounded continuous function. The spectral measure formula is thus obtained, using the continuity of the action of a measure on bounded continuous function and the unicity of the spectral measure, by going to the limit in the previous equality.

The $h \in \ell^2$ case: If we assume that $\mathbb{E}X[0] = 0$ and $R_X \in \ell^1$ then $Y = h \star X$ exists even if $h \in \ell^2$ and is a WSSP of autocovariance function

$$R_Y = h \star \tilde{h} \star R_X,$$

where $\tilde{h}$ is defined $\forall n \in \mathbb{Z}, \tilde{h}[n] = \overline{h[-n]}$. Moreover, the spectral measure μ_Y of Y is given by $d\mu_Y(\omega) = |\hat{h}(e^{i\omega})|^2 d\mu_X(\omega) = |\hat{h}(e^{i\omega})|^2 \widehat{R_X}(e^{i\omega}) \mathrm{d}\omega$.

Proof: The proof is again based on $Y_N = h_N \star X$ which is a well defined zero mean stationary process. Now $Y_N - Y_M = (h_N - h_M) \star X$ is also well defined and satisfies

$$R_{Y_N - Y_M} = (h_N - h_M) \star \widetilde{(h_N - h_M)} \star R_X.$$

If we use the property that $\|g \star g'\|_\infty \leq \min(\|g\|_2 \|g\|_2, \|g\|_\infty \|g'\|_1$, one obtains thus

$$\begin{aligned} \|R_{Y_N - Y_M}\|_\infty &\leq \|h_N - h_M\|_2 \|\widetilde{(h_N - h_M)}\|_2 \|R_X\|_1 \\ &\leq \|R_X\|_1 \|h_N - h_M\|_2^2. \end{aligned}$$

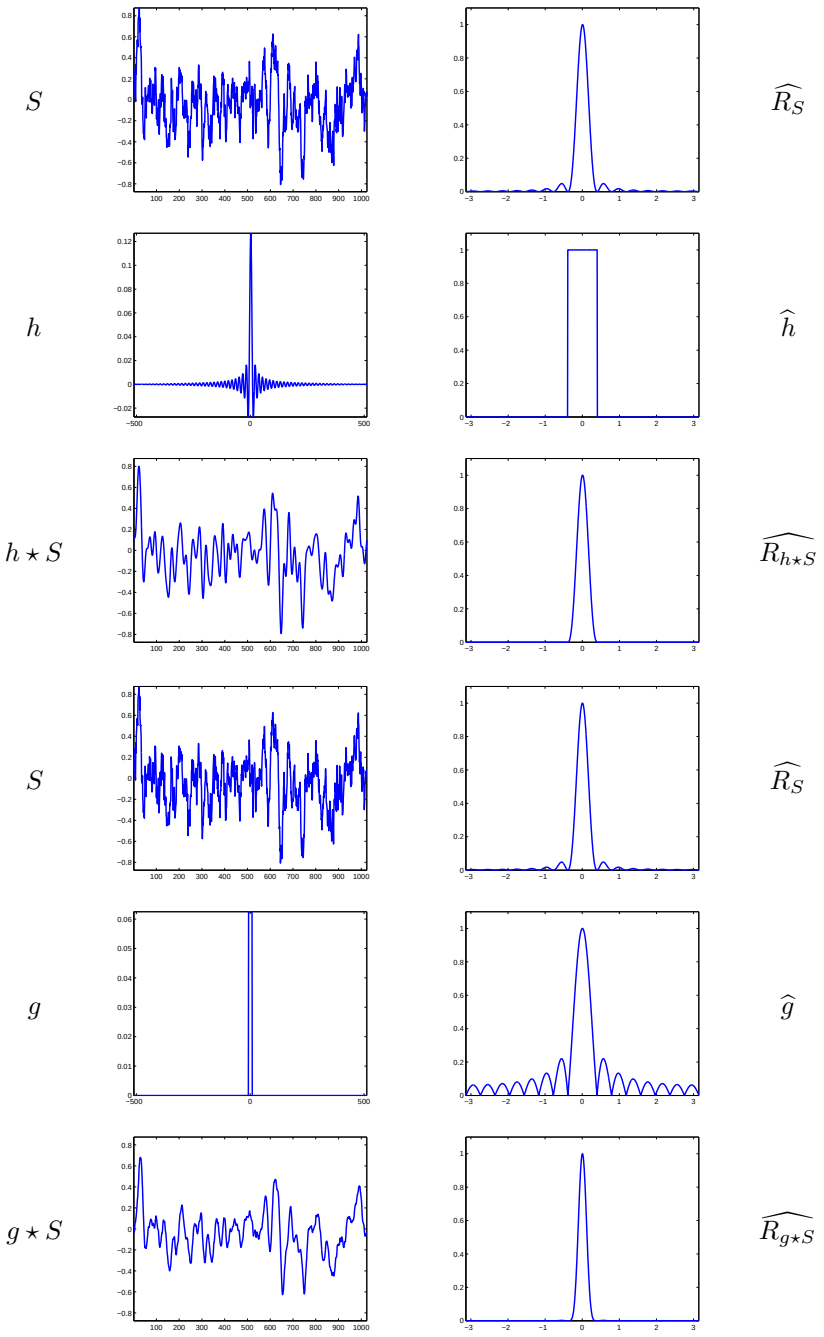


Figure 4.6: Process filtering

As $h \in L^2$, h_N is a Cauchy sequence in L^2 and thus $\forall \epsilon > 0$ it exists N_0 such that $N \geq N_0$ and $M \geq N_0$ implies $\|R_{Y_N - Y_M}\|_\infty \leq \epsilon$. As $\mathbb{E}Y_N[n] = 0$, this implies that $Y_N[n]$ has a limit $Y[n]$ for the convergence in L^2 and that this limits satisfies $\mathbb{E}Y[n] = 0$. Now using the continuity of the convolution with respect to the ℓ^2 norm of the filter on verify that

$$R_{Y_N} = h_N \star \widetilde{h_N} \star R_X \to h \star \widetilde{h} \star R_X = R_Y.$$

Finally,

$$R_{Y_N}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{in\omega} \left|\widehat{h_N}(e^{i\omega})\right|^2 \widehat{R_X}(e^{i\omega}) \mathrm{d}\omega$$

which converge to

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{in\omega} \left|\widehat{h}(e^{i\omega})\right|^2 \widehat{R_X}(e^{i\omega}) \mathrm{d}\omega$$

because $R_X \in \ell^1$ implies $\widehat{R_X} \in L^\infty$ and $h \in \ell^2$ implies that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \left|\widehat{h_N}(e^{i\omega})\right|^2 \mathrm{d}\omega \to \frac{1}{2\pi} \int_{-\pi}^{\pi} \left|\widehat{h}(e^{i\omega})\right|^2 \mathrm{d}\omega.$$

Remark: If X has a spectral measure of density $\hat{R}(e^{i\omega})$ with respect to the Lebesgue measure which is continuous, the filtering of X by the filter h_{ϵ,ω_0} whose Fourier transform is given by

$$\widehat{h_{\omega_0,\epsilon}}(e^{i\omega}) = \frac{1}{\sqrt{\epsilon}} \mathbf{1}_{\{|\omega_0 - \omega| < \epsilon/2\}}$$

yields a centered random process Y_{ϵ,ω_0} whose variance tends to $\hat{R}(e^{i\omega_0})$ when ϵ tends to 0. The name of spectral density (or power spectral density) is due to this.

4.2 Wiener filtering

4.2.1 Filter to estimate?

In many applications, instead of observing a WSSP X, one observes a WSSP D linked to X through its covariance and one wishes to estimate X from D. This model is rather general and includes for instance the cases of

- denoising: $D = X + B$ where B is a independent WSSP of *noise*,
- deconvolution: $D = g \star X + B$ where g is a know filter.

All signals being WSSP, it is natural to use a translation invariant method. If one further imposes that the method is linear, one imposes to estimate X from D by a LTI system and thus by the convolution of D with a filter h.

4.2.2 Wiener filtering

Theorem 4.4: Let X be a WSSP of zero mean and D another WSSP of zero mean such that the covariance function

$$R_{XD}[n,m] = \mathrm{Cov}(X[n], D[m])$$

depends only on $n-m$. If it exists $h \in \ell^1$ such that

$$R_{XD} = h \star R_D$$

then the best linear prediction of X in the space generated by D in term of quadratic error is given by

$$\tilde{X} = h \star D$$

and

$$\mathbb{E}\left(|\tilde{X}[n] - X[n]|^2\right) = R_X[0] - \sum_{k\in\mathbb{Z}} \overline{h[k]} R_{XD}[k]$$

Proof: Let $n \in \mathbb{Z}$, the best prediction $\tilde{X}[n]$, in term of quadratic error, of $X[n]$ in the space generated by all $D[k]$ with $k \in \mathbb{Z}$ is by the hermitian structure of the L^2 space, the projection of $X[n]$ onto this subspace. This projection is entirely characterized by the projection theorem:

- $\tilde{X}[n] \in \mathrm{Vect}(D)$
- and $\forall k \in \mathbb{Z}, \mathrm{Cov}(X[n] - \tilde{X}[n], D[k]) = 0$.

Moreover

$$\begin{aligned}\mathbb{E}(|X[n] - \tilde{X}[n]|^2) &= \mathrm{Cov}(X[n] - \tilde{X}[n], X[n] - \tilde{X}[n]) = \mathrm{Cov}(X[n], X[n] - \tilde{X}[n]) \\ &= \mathrm{Cov}(X[n], X[n]) - \mathrm{Cov}(X[n], \tilde{X}[n]).\end{aligned}$$

Let $h \in \ell^1$ such that

$$R_{XD} = h \star R_D,$$

the process $Z = h \star D$ is a WSSP defined $\forall n \in \mathbb{Z}$ by

$$Z[n] = \sum_{k\in\mathbb{Z}} h[k] D[n-k].$$

By construction, $Z[n] \in \mathrm{Vect}(D)$. Furthermore, $\forall (n,k) \in \mathbb{Z}$,

$$\begin{aligned}\mathrm{Cov}(Z[n], D[k]) &= \sum_{l\in\mathbb{Z}} h[l] \,\mathrm{Cov}(D[n-l], D[k]) = \sum_{l\in\mathbb{Z}} h[l] R_D[n-l-k] \\ &= (h \star R_D)[n-k].\end{aligned}$$

We have thus

$$\begin{aligned}\mathrm{Cov}(X[n] - Z[n], D[k]) &= \mathrm{Cov}(X[n], D[k]) - \mathrm{Cov}(Z[n], D[k]) \\ &= R_{XD}[n-k] - (h \star R_D)[n-k] \\ &= 0.\end{aligned}$$

This implies that $Z[n]$ is the projection of $X[n]$ on $\mathrm{Vect}(D)$ and thus $\tilde{X}[n] = Z[n] = (h \star D)[n]$.

The last item of the theorem comes from

$$\begin{aligned}\mathbb{E}(|X[n]-\tilde{X}[n]|^2) &= \mathrm{Cov}(X[n],X[n]) - \mathrm{Cov}(X[n],\tilde{X}[n]) \\ &= \mathrm{Cov}(X[n],X[n]) - \mathrm{Cov}(X[n],\sum_{k\in\mathbb{Z}} h[k]D[n-k]) \\ &= \mathrm{Cov}(X[n],X[n]) - \sum_{k\in\mathbb{Z}} \overline{h[k]}\,\mathrm{Cov}(X[n],D[n-k]) \\ \mathbb{E}(|X[n]-\tilde{X}[n]|^2) &= R_X[0] - \sum_{k\in\mathbb{Z}} \overline{h[k]} R_{XD}[k].\end{aligned}$$

Corollary 4.3: Under the assumptions of the previous theorem, if $R_X \in \ell^1$, $R_D \in \ell^1$ and $R_{XD} \in \ell^1$, $\hat{h}$ satisfies

$$\hat{h}(e^{i\omega}) = \frac{\hat{R}_{XD}(e^{i\omega})}{\hat{R}_D(e^{i\omega})}$$

and

$$\mathbb{E}\left(|X[n]-\tilde{X}[n]|^2\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\hat{R}_X(e^{i\omega}) - \frac{|\hat{R}_{XD}(e^{i\omega})|^2}{\hat{R}_D(e^{i\omega})}\right)d\omega.$$

If $\hat{R}_D(e^{i\omega})$ in non zero when $\hat{R}_X(e^{i\omega})$ is non zero

$$\mathbb{E}\left(|X[n]-\tilde{X}[n]|^2\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\hat{R}_X(e^{i\omega})\hat{R}_D(e^{i\omega}) - |\hat{R}_{XD}(e^{i\omega})|^2}{\hat{R}_D(e^{i\omega})}d\omega.$$

Proof: The first item is obtained by computing the Fourier transform of the equality $R_{XD} = h \star R_D$ which involves ℓ^1 discrete functions and thus yields almost surely

$$\hat{R}_{XD}(e^{i\omega}) = \hat{h}(e^{i\omega})\hat{R}_D(e^{i\omega}).$$

This implies that one can define $\hat{h} = \hat{R}_{XD}(e^{i\omega})/\hat{R}_D(e^{i\omega})$ when $R_D(e^{i\omega}) \neq 0$ and otherwise arbitrarily because if $\hat{R}_D(e^{i\omega}) = 0$ then $\hat{R}_{XD}(e^{i\omega}) = 0$.

The second item is obtained from

$$\mathbb{E}\left(|\tilde{X}[n]-X[n]|^2\right) = R_X[0] - \sum_{k\in\mathbb{Z}} \overline{h[k]} R_{XD}[k]$$

which becomes using the spectral density of R_X et R_{XD} which are both ℓ^1, the reconstruction formula and Parseval equality

$$\begin{aligned}&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\hat{R}_X(e^{i\omega})d\omega - \frac{1}{2\pi}\int_{-\pi}^{\pi}\overline{\hat{h}(e^{i\omega})}\hat{R}_{XD}(e^{i\omega})d\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\hat{R}_X(e^{i\omega}) - \frac{|\hat{R}_{XD}(e^{i\omega})|^2}{\hat{R}_D(e^{i\omega})}\right)d\omega\end{aligned}$$

and thus because $\hat{R}_D(e^{i\omega})$ does not vanish where $\hat{R}_X(e^{i\omega})$ is non zero

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\hat{R}_X(e^{i\omega})\hat{R}_D(e^{i\omega}) - |\hat{R}_{XD}(e^{i\omega})|^2}{\hat{R}_D(e^{i\omega})}d\omega.$$

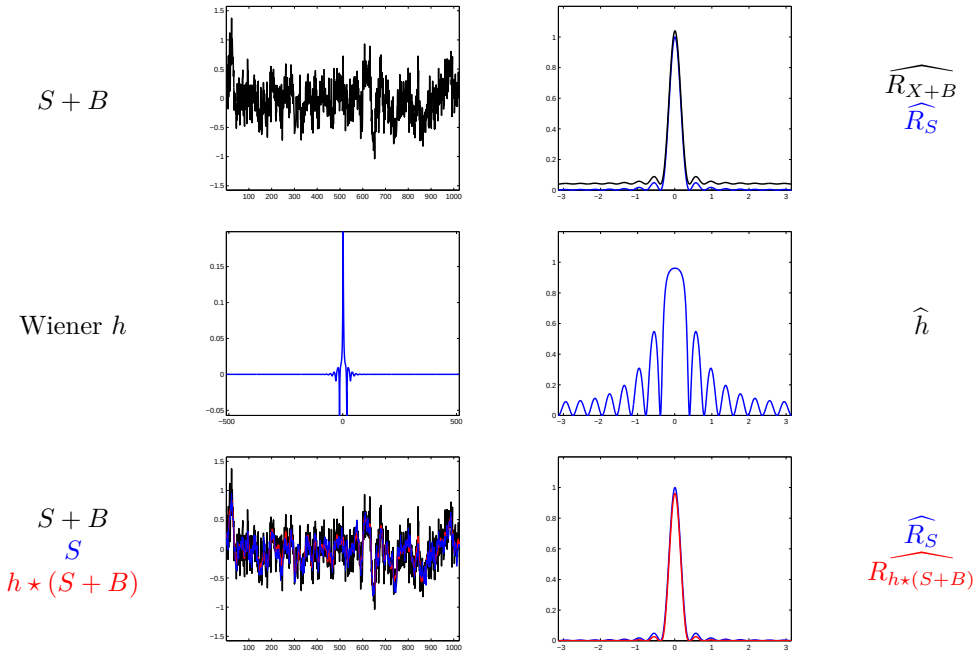


Figure 4.7: Wiener filtering

Examples: Let X and B be two independent zero mean WSSP with known spectral densities.

- Denoising: one observes $D = X + B$, then

$$\hat{R}_D(e^{i\omega}) = \hat{R}_X(e^{i\omega}) + \hat{R}_B(e^{i\omega}) \quad \text{and} \quad \hat{R}_{XD}(e^{i\omega}) = \hat{R}_X(e^{i\omega})$$

  and if $\exists h \in \ell^1$ such that $R_D = h \star R_{XD}$ then

$$\hat{h} = \frac{\hat{R}_X(e^{i\omega})}{\hat{R}(e^{i\omega})_X + \hat{R}_B(e^{i\omega})} \quad \text{and} \quad \mathbb{E}\left(|X[n] - \tilde{X}[n]|^2\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\hat{R}_B(e^{i\omega})}{\hat{R}_X(e^{i\omega}) + \hat{R}_B(e^{i\omega})} \hat{R}_X(e^{i\omega}) d\omega.$$

- Deconvolution: on observes $D = g \star X + B$ with $g \in \ell^1$ known, then

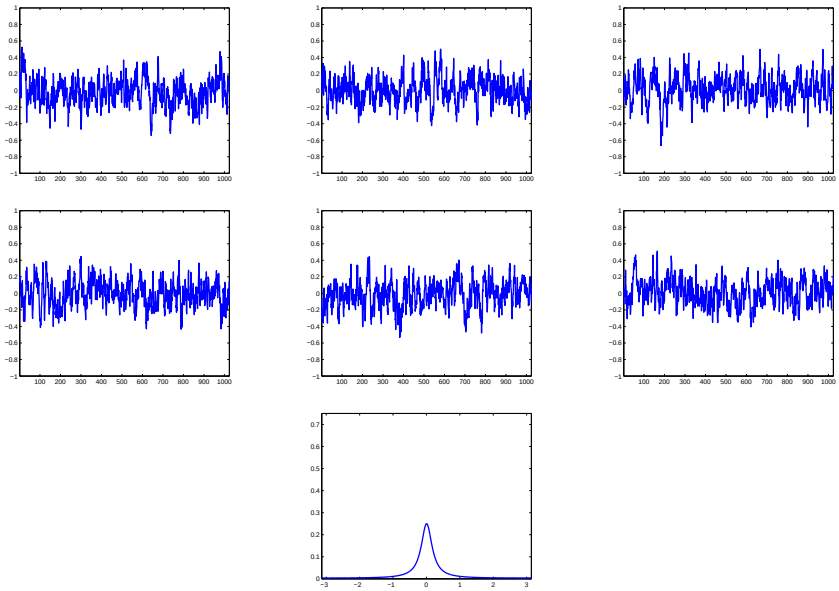
$$\hat{R}_D(e^{i\omega}) = |\hat{g}(e^{i\omega})|^2 \hat{R}_X(e^{i\omega}) + \hat{R}_B(e^{i\omega}) \quad \text{and} \quad \hat{R}_{XD}(e^{i\omega}) = \overline{\hat{g}}(e^{i\omega}) \hat{R}_X(e^{i\omega})$$

  and if $\exists h \in \ell^1$ such that $R_D = h \star R_{XD}$ then

$$\hat{h} = \frac{\overline{\hat{g}}(e^{i\omega}) \hat{R}_X(e^{i\omega})}{|\hat{g}(e^{i\omega})|^2 \hat{R}(e^{i\omega})_X + \hat{R}_B(e^{i\omega})} \quad \text{and} \quad \mathbb{E}\left(|X[n] - \tilde{X}[n]|^2\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\hat{R}_B(e^{i\omega})}{|\hat{g}(e^{i\omega})|^2 \hat{R}_X(e^{i\omega}) + \hat{R}_B(e^{i\omega})} \hat{R}_X(e^{i\omega}) d\omega.$$

In practice, Wiener filtering method consists in computing the inverse Fourier transform of $\widehat{h} = \frac{\hat{R}_{XD}(e^{i\omega})}{\hat{R}_D(e^{i\omega})}$ and checking that it belongs to ℓ^1. If it is the case, the theorem applies otherwise one has to try to adapt the proof. For instance, if $\widehat{h} \in L^2([-\pi, \pi])$ then $h \in \ell^2$ and one can still use this filter if R_X, R_{XD} and R_D are in ℓ^1.

$$X[n] = -\frac{p_1}{p_0}X[n-1] - \frac{p_2}{p_0}X[n-2] + \frac{1}{p_0}B[n]$$



$$\widehat{R_X} = \frac{1}{|p_0 - p_1 e^{-i\omega} - p_2 e^{-2i\omega}|^2}$$

Figure 4.8: AR example

4.3 ARMA process

4.3.1 AR process and canonical decomposition

The AR processes are simple processes corresponding to recursive sequence in the deterministic case. They are frequently used in random modeling because they are both simple and versatile.

Definition: A WSSP X is autoregessive (AR) if it exists a non zero polynomial P and a normalized white noise B such that

$$p \star X = B$$

where p is the filter associated to P ($\hat{p}(e^{i\omega}) = P(e^{-i\omega})$).

Provided $P_0 = P(0) \neq 0$, we have thus a recursive definition of $X[n]$

$$p_0 X[n] = -\sum_{k=1}^{K} p_k X[n-k] + B[n]$$

P cannot be arbitrary.

Proposition 4.4: P has no roots of modulus 1.

Proof: By the filtering theorem, $|\hat{p}(e^{i\omega})|^2 \mathrm{d}\mu_X(\omega) = \mathrm{d}\mu_B(\omega)$ i.e. $|P(e^{-iw})|\mathrm{d}\mu_X(\omega) = d\omega$.

Mean: In particular, as $\mathbb{E}p \star X = P(1)\mathbb{E}X = E(B) = 0$, this implies that $\mathbb{E}X[0] = 0$.

Inverse filter: If p has an inverse $p^{-1} \in \ell^1$ then $X = p^{-1} \star B$.

Proposition 4.5: If P is a polynomial without roots of modulus 1 such that $P(0) \neq 0$ then its associated filter p admits a stable inverse p^{-1}. Furthermore, this filter is causal if and only if the roots of P are of modulus strictly greater than 1. Lastly, $p^1[0] = P(0)^{-1}$.

Proof: The proof is similar to the one of the recursive filter inversion and is based on the partial fraction decomposition of $1/P(X)$.

Stable and causal inverse: as it was the case for recursive filtering, one can always assume this is the case by modifying P.

Proposition 4.6: Let P be a polynomial without roots of modulus 1 and such that $P(0) \neq 0$, it exists a unique polynomial P_0 and a unique positive real σ_0 such that

- $\forall z \in \mathbb{C}|z| = 1 \implies \sigma_0|P(z)| = |P_0(z)|$
- every root of P_0 has a modulus strictly greater than 1
- $P_0(0) = 1$

Canonical couple: The couple (P_0, σ_0) associated to P is called the canonical couple associated to P.

Proof: Existence:

$$|P(z)| = \left| c \prod_{i=1}^{N} (z - \xi_i) \right|$$

where the ξ_i are the roots of P

$$= |c| \prod_{i=1}^{N} |z - \xi_i|.$$

One verify easily that for all $z \in \mathbb{C}$ of modulus 1 and all $\xi \in \mathbb{C}$,

$$|z - \xi| = |\overline{z}||z - \xi| = |1 - \xi\overline{z}| = |\overline{\xi}z - 1| = |\xi| \left| z - \frac{1}{\overline{\xi}} \right|.$$

One can thus replace every occurrence of $|z-\xi|$ by $|\xi_i| \left| z - \frac{1}{\overline{\xi_i}} \right|$ for all roots ξ_i of modulus strictly smaller than 1. Up to a reordering of the roots, this require to modify only the N' first roots, so that

$$\begin{aligned} |P(z)| &= |c| \prod_{i=1}^{N'} |\xi_i| \left| z - \frac{1}{\overline{\xi_i}} \right| \prod_{i=N'+1} N|z - \xi_i| \\ &= \left(|c| \prod_{i=1}^{N'} |\xi_i| \right) \left| \prod_{i=1}^{N'} \left(z - \frac{1}{\overline{\xi_i}} \right) \prod_{i=N'+1}^{N} (z - \xi_i) \right| \\ &= \frac{1}{\sigma_0} |P_0(z)| \end{aligned}$$

Uniqueness: Let A and B two polynomials such that $|A(z)| = c|B(z)|$ on the unit circle, $A(0) = B(0) = 1$ and such that their roots have a modulus > 1.

Let ξ_k be the roots of A andη_k the ones of B: for z of modulus 1,

$$\prod_{i=1}^{N}(z-\xi_i)(\overline{z}-\overline{\xi_i}) = c^2\prod_{i=1}^{M}(z-\eta_i)(\overline{z}-\overline{\eta_i})$$

Multiplying on the left and on the right by z^{N+M} and using $z\overline{z} = 1$ yields

$$z^M\prod_{i=1}^{N}(z-\xi_i)(1-\overline{\xi_i}z) = c^2z^N\prod_{i=1}^{M}(z-\eta_i)(1-\overline{\eta_i}z).$$

Those two polynomials are equal on the unit circle and thus equals. The set of their roots are thus identical and

$$\{\xi_k, \frac{1}{\overline{\xi_k}}, 1 \leq k \leq N\} = \{\eta_k, \frac{1}{\overline{\eta_k}}, 1 \leq k \leq M\}$$

This implies that $N = M$ and the assumptions $|\xi_k| > 1$ and $\eta_k| > 1$ ensure that

$$\{\xi_k, 1 \leq k \leq N\} = \{\eta_k, 1 \leq k \leq M\}.$$

The two polynomials A and B are thus proportional and as $A(0) = B(0)$ are identical.

Theorem 4.5: Let X be an AR WSSP, it exists a unique couple (P_0, σ_0) such that

- the filter p_0 associated to P_0 satisfies$p_0 \star X = \sigma_0 B_0$ where B_0 is a normalized white noise.
- $P_0(0) = 1$ and all the roots of P_0 have a modulus > 1.

Such a couple is called the canonical decomposition of the AR WSSP X.

Corollary 4.4: if X is an AR WSSP, it exist a unique stable causal filter p_0^{-1}, a unique standard deviation σ_0 and a normalize white noise B_0 such that

$$X = \sigma_0 p_0^{-1} \star B_0$$

Proof: X is an AR WSSP and thus it exists a polynomial P such that the filter p associated to X satisfies $p \star X = B$ with B a normalized white noise. Let (P_0, σ_0) be the canonical couple associated to P. The process $p_0 \star X$ is a WSSP process of power spectral density

$$|\hat{p}_0(e^{i\omega})|^2 \mathrm{d}\mu_X(\omega) = \sigma^2|\hat{p}(e^{i\omega})|^2 \mathrm{d}\mu_X(\omega) = \sigma_0^2 d\omega.$$

and thus $\frac{1}{\sigma_0}p_0 \star X$ is a normalized white noise.

Let (P_1, σ_1) be a couple satisfying the same constraints, we have thus

$$\frac{1}{\sigma^2}|\hat{p}_0(e^{i\omega})|^2 \mathrm{d}\mu_X(\omega) = \frac{1}{\sigma_1^2}|\hat{p}_1(e^{i\omega})|^2 \mathrm{d}\mu_X(\omega)$$

which implies

$$|P_0(e^{-i\omega})|^2 \mathrm{d}\mu_X(\omega) = \frac{\sigma_0^2}{\sigma_1^2}|P_1(e^{-i\omega})|^2 \mathrm{d}\mu_X(\omega).$$

This implies thus that $P_0 = P_1$ and $\sigma_0 = \sigma_1$.

4.3.2 Prediction and coefficient estimation

Prediction: How to predict $X[n+1]$ from the already observed values of X: $X[n-k]$ for $k \in \mathbb{N}$?

Projection: If one restricts ourself to linear prediction, the very same analysis than the one conducted to study Wiener filtering shows that the best prediction $\tilde{X}[n+1]$ is the projection of $X[n+1]$ onto the space $\mathrm{Vect}_n(X) = \mathrm{Vect}(X[n-k], k \in \mathbb{N})$ that we denote $P_{\mathrm{Vect}_n(X)}(X[n+1])$.

Theorem 4.6: Let X be AR WSSP of canonical decomposition $p_0 \star X = \sigma_0 B_0$ then

$$P_{\mathrm{Vect}_n(X)}(X[n+1]) = -\sum_{k>0} p_0[k+1]X[n-k]$$

and

$$\mathbb{E}\left(\left|X[n+1] - P_{\mathrm{Vect}_n(X)}(X[n+1])\right|^2\right) = \sigma_0^2.$$

The proof relies heavily on the following proposition:

Proposition 4.7: If X is an AR WSSP with canonical decomposition $p_0 \star X = \sigma_0 B_0$ then

- the filter p_0^{-1} is a stable causal filter such that
  - $X = \sigma_0 p_0^{-1} \star B_0$ where B_0 is a normalized white noise
  - $p_0^{-1}[0] = 1$.
- $\mathrm{Vect}_n(X) = \mathrm{Vect}_n(B_0)$.

Proof: The first part is a restatement of a previous proposition.

The second part is obtained by noticing that the canonical decomposition yields

$$B_0[n] = \frac{1}{\sigma_0}(p_0 \star X)[n] = \frac{1}{\sigma_0}\sum_{k\geq 0} p_0 n[k]X[n-k]$$

because p_0 is a stable causal filter. This implies that $\mathrm{Vect}_n(B_0) \subset \mathrm{Vect}_n(X)$. Along the same line,

$$X[n] = \sigma_0(p_0^{-1} \star B_0)[n] = \sigma_0 \sum_{k\geq 0} p_0^{-1}[k]B_0[n-k]$$

because p_0^{-1} is also a stable causal filter. This implies that $\mathrm{Vect}_n(X) \subset \mathrm{Vect}_n(B_0)$.

Proof (of the theorem): As $\mathrm{Vect}_n(X) = \mathrm{Vect}_n(B_0)$ and B_0 is a white noise $B_0[n+1]$ is uncorelated with $\mathrm{Vect}_n(X)$. This can be rewritten as

$$P_{\mathrm{Vect}_n(X)}(B_0[n+1]) = 0.$$

The canonical decomposition of X at $n+1$ yields

$$\sigma_0 B_0[n+1] = X[n+1] + \sum_{k>0} p_0[k+1]X[n-k]$$

and thus if we project this equality

$$\sigma_0 P_{\mathrm{Vect}_n(X)}(B_0[n+1]) = P_{\mathrm{Vect}_n(X)}(X[n+1]) + \sum_{k\geq 0} p_0[k+1] P_{\mathrm{Vect}_n(X)}(X[n-k])$$

which yields

$$0 = P_{\mathrm{Vect}_n(X)}(X[n+1]) + \sum_{k>0} p_0[k+1] X[n-k]$$

which is the first item.

The second one is immediate once we notice that $X[n+1] - P_{\mathrm{Vect}_n(X)}(X[n-k]) = \sigma_0 B_0$.

Corollary 4.5: The best linear prediction of $X[n+k]$ fromy $\mathrm{Vect}_n(X)$, $P_{\mathrm{Vect}_n(X)}(X[n+k])$ is obtained by applying the previous formula recursively

$$P_{\mathrm{Vect}_n(X)}(X[n+k]) = P_{\mathrm{Vect}_{n+k-1}(X)} \cdot P_{\mathrm{Vect}_n(X)}(X[n+k])$$

and

$$\mathbb{E}\left(\left|X[n+k] - P_{\mathrm{Vect}_n(X)}(X[n+k])\right|^2\right) = \sigma_0^2 \sum_{l=0}^{k-1} |p_0^{-1}[l]|^2.$$

Proof: The first item is immediate as $\mathrm{Vect}_{n'}(X) \subset \mathrm{Vect}_{n'+1}(X)$. For the second one, starting from the decomposition

$$X[n+l] = \sigma_0 \sum_{l\geq 0} p_0^{-1}[l] B_0[n+k-l]$$

and projecting it on $\mathrm{Vect}_n(X) = \mathrm{Vect}_n(B_0)$ yields

$$P_{\mathrm{Vect}_n(X)}(X[n+l]) = \sigma_0 \sum_{l\geq k} p_0^{-1}[l] B_0[n+k-l]$$

which implies

$$X[n+l] - P_{\mathrm{Vect}_n(X)}(X[n+l]) = \sigma_0 \sum_{l=0}^{k-1} p_0^{-1}[l] B_0[n+k-l]$$

which yields the result because B_0 is normalized white noise.

Estimation: In practice, the parameters of the AR WSSP are often not known and need to be estimated.

Using the representation $X = \sigma_0 p_0^{-1} \star B_0$ one can show that both the covariance estimates $\forall 0 \leq k \leq N-1$

$$\widetilde{R}_X[k] = \frac{1}{N} \sum_{l=0}^{N-1-k} (X[n] - \widetilde{\mu})\overline{(X[n-k] - \widetilde{\mu}}$$

are consistant for an AR process.

Now as

$$p_0 \star X[N] = \sigma B_0[N]$$

we have for all k'

$$\sum_{k=0}^{K} p_0[k] \operatorname{Cov}(X[N-k], X[N-k']) = \sigma \operatorname{Cov}(B_0[N], X[N-k'])$$

and thus

$$\sum_{k=0}^{K} p_0[k] R_X[k'-k] = \sigma^2 \delta_{k'=0}$$

Pluging our estimate of R_X for $0 \leq k' \leq K$ allows to estimate p_0 and σ^2 from this set of linear equation, called the Yule-Walker system.

An efficient implementation of the resolution of this system known as the Levinson-Durbin is often used in practice instead of a brute force inversion.

Yule-Walker as a least square:

Theorem 4.7: The minimizer in p_0 of

$$\sum_{n=K}^{N-1} \left| X[n] + \sum_{k=1}^{K} p_0[k] X[n-k] \right|^2$$

is the solution of the Yule-Walker system.

Proof: A necessary condition to minimize the convex function

$$\sum_{n=K}^{N-1} \left| X[n] + \sum_{k=1}^{K} p_0[k] X[n-k] \right|^2,$$

is that the gradient at that point is 0 and thus that for all $1 \leq k' \leq K$

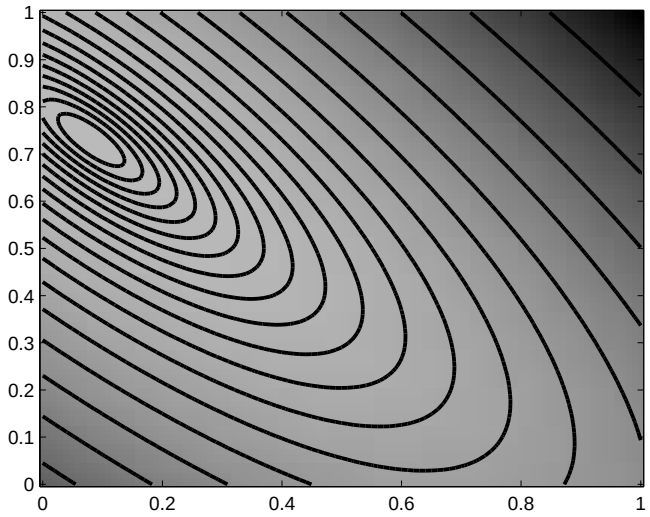
$$\begin{aligned} 2 \sum_{n=K}^{N-1} \sum k = 0^K p_0[k] X[n-k] \overline{X[n-k']} &= 0 \\ \sum_{k=0}^{K} p_0[k] \sum_{n=K}^{N+1} X[n-k] \overline{X[n-k']} &= 0 \\ \sum_{k=0}^{K} p_0[k] N \widetilde{R}_X[k'-k] &= 0. \end{aligned}$$

4.3.3 Generalization to ARMA process

Definition: A WSSP X is an ARMA (Auto Regressive Moving Average) if there exist a non zero polynomial P, a non zero polynomial Q without roots of modulus 1 and a normalized white noise B such that

$$p \star X = q \star B$$

where p (respectively q) is the filter associated to P (respectively Q).



$$\frac{1}{n}\sum_{n=3}^{N}|X[n]-\alpha_1 X[N-1]-\alpha_2 X[N-2]|^2$$

Figure 4.9: AR parameter estimation

Theorem 4.8: Let X be an ARMA WSSP, it exist a unique triplet (P_0, Q_0, σ_0) such that

- $p_0 \star X = \sigma_0 q_0 \star B_0$ where B_0 is a normalized white noise.
- $P_0(0) = Q_0(0) = 1$ and the roots of P_0 and Q_0 have a modulus > 1.
- P_0 and Q_0 no common roots.
- p_0 admit a stable causal inverse p_0^{-1} verifying $p_0^{-1}[0] = 1$ such that $X = \sigma_0 p_0^{-1} \star q_0 \star B_0$.
- q_0 admit a stable causal inverse q_0^{-1} verifying $q_0^{-1}[0] = 1$ such that $\sigma B_0 = q_0^{-1} \star p_0 \star X$.

Theorem 4.9: If X is a ARMA WSSP with canonical decomposition $p_0 \star X = \sigma_0 q_0 \star B_0$ then $\mathrm{Vect}_n(X) = \mathrm{Vect}_n(B_0)$,

$$P_{\mathrm{Vect}_n(X)}(X[n+1]) = -\sum_{k\geq 0}(q_0^{-1} \star p_0)[k+1]X[n-k]$$

and

$$\mathbb{E}\left(\left|X[n+1] - P_{\mathrm{Vect}_n(X)}(X[n+1])\right|^2\right) = \sigma_0^2$$

Corollary 4.6: The best linear prediction of $X[n+k]$ from $\mathrm{Vect}_n(X)$, is obtained by applying the previous formula recursively

$$P_{\mathrm{Vect}_n(X)}(X[n+k]) = P_{\mathrm{Vect}_{n+k-1}(X)} \cdot P_{\mathrm{Vect}_n(X)}(X[n+k])$$

and

$$\mathbb{E}\left(\left|X[n+k] - P_{\mathrm{Vect}_n(X)}(X[n+k])\right|^2\right) = \sigma_0^2 \sum_{l=0}^{k-1} |(p_0^{-1} \star q_0)[l]|^2.$$

Coefficient estimation: For ARMA process, using the representation $X = \sigma_0 p_0^{-1} \star q_0 \star B_0$ one can show that all the covariance estimates $\forall 0 \leq k \leq N-1$

$$\widetilde{R}_X[k] = \frac{1}{N} \sum_{l=0}^{N-1-k} (X[n] - \widetilde{\mu})\overline{(X[n+k] - \widetilde{\mu}}$$

are consistent. This is the basis of most methods of estimation of the parameters.

Chapter 5

Voice modeling

In this chapter, we study the voice production mechanism. We start by a physiological description that leads to a physical model. We show how to compute solutions from this model using Fourier technique and how this modeling can be used to synthesize efficiently voices.

5.1 Anatomy and physics

5.1.1 Anatomical description

The voice production mechanism can be divided in three phases:

- **Respiration:** Air goes from the lungs to the trachea
- **Phonation:** In the larynx, with the help of the vocals cords, an *excitation* signal is produced.
- **Articulation:** This signal is deformed by the vocal tract to produce an *articulated* sound.

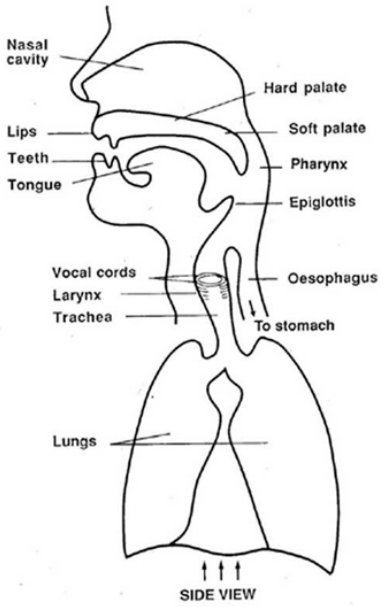
In the phonation step, the vocal cords can either vibrate to produce a *quasi* periodic signal (voiced sound) or let the air go through to produce a turbulent signal that can be modeled by a white noise (unvoiced sound). The frequency of vibration of the vocal cords is called the *pitch* and gives the fundamental frequency of the sound. This pitch is typically between 100 Hz and 300 Hz, but can be as high as 3000 Hz for a soprano.

In the articulation phase, the signal produced by the vocal folds in *filtered* by the vocal tract to produce *articulated* sound. Playing with the shape of the mouth, pharynx and nasal cavities, one obtains different sounds.

5.1.2 Physical modeling

Physics of tube: we focus on the effect of the vocal tract which one models by a simple 1D tube. The physical behavior will be described by four quantities:

X-ray view



scheme

Figure 5.1: Voice anatomy

Figure 5.2: Vocal cords

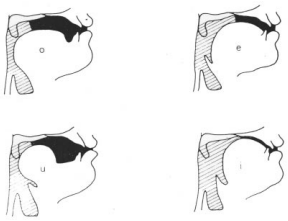


Figure 5.3: Vocal tract for different vowels

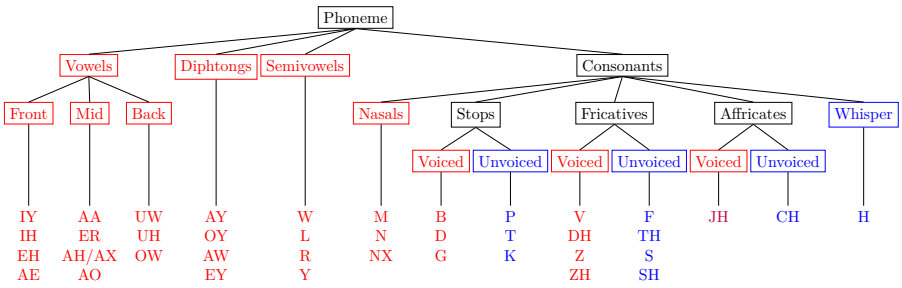


Figure 5.4: Phoneme classification

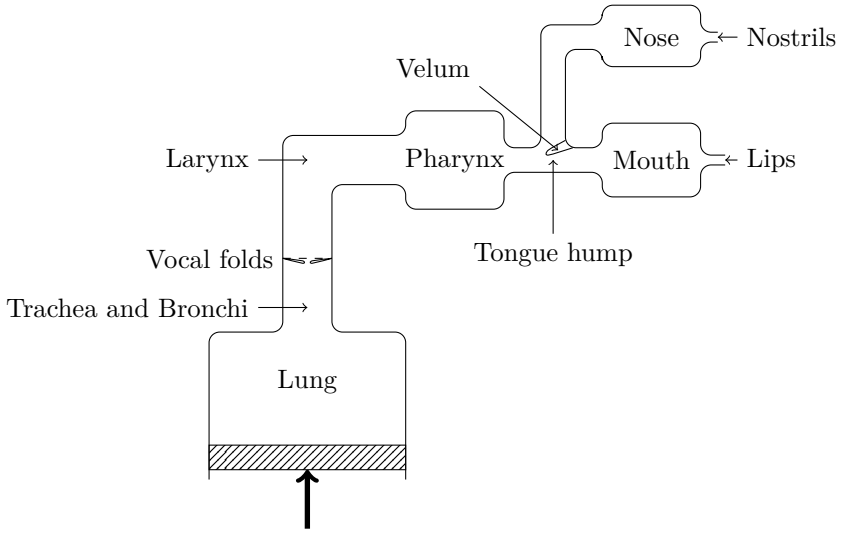


Figure 5.5: Speech production system

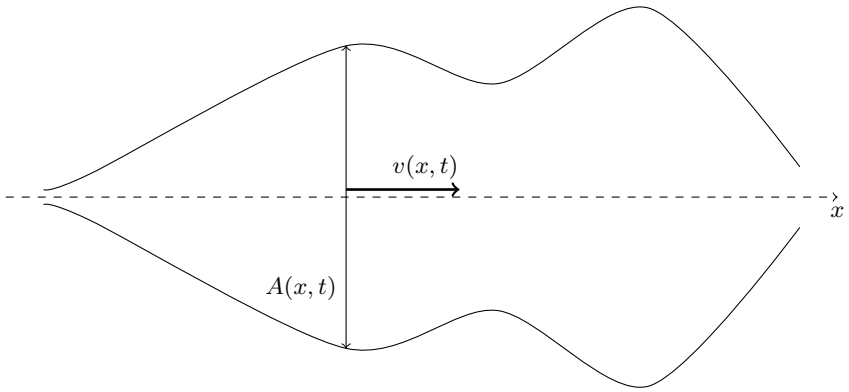


Figure 5.6: Tube model

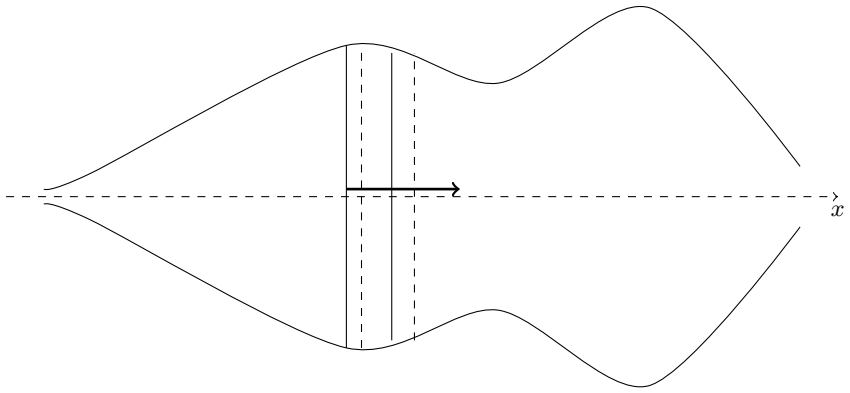


Figure 5.7: Mass conservation

- the tube area $A(x,t)$
- the air velocity: $v(x,t)$
- the Pressure: $P(x,t)$
- and the air density: $\rho(x,t)$.

Pressure and density: in an adiabatic gas, one has $PV^\gamma = \text{cst}$ or equivalently $P\rho^{-\gamma} = \text{cst}'$. We will assume here the more general relationship $P = f(\rho)$ where f increases with ρ. Physical consideration shows that the air density can not vary drastically from the ambient air density so that $\rho = \rho_0 + \rho_e$ with $\rho_e \ll \rho_0$ (Ambient air density). We can thus use a Taylor expansion to obtain (at least approximately) that

$$P = f(\rho_0) + \kappa^2 \rho_e$$

with $\kappa^2 = \frac{\partial f}{\partial \rho}(\rho_0)$. In particular

$$\mathrm{d}P = \kappa^2 \mathrm{d}\rho$$

Mass conservation: one of the most fundamental principle of physics is mass conservation. The mass of an infinitesimal slide at (x,t) is given by

$$\rho(x,t)A(x,t)\mathrm{d}x$$

while the mass of the same slide at time $t + \mathrm{d}t$ is given by

$$\begin{aligned}&\rho(x + v(x,t)\mathrm{d}t, t + \mathrm{d}t)A(x + v(x,t)\mathrm{d}t, t + \mathrm{d}t)\\ &\times [x + \mathrm{d}x + v(x + \mathrm{d}x, t)\mathrm{d}t - (x + v(x,t)\mathrm{d}t)]\end{aligned}$$

So gathering ρ and A, we obtain

$$\rho A \mathrm{d}x = \left(\rho A + \frac{\partial \rho A}{\partial x} v \mathrm{d}t + \frac{\partial \rho A}{\partial t}\mathrm{d}t\right)\left(\mathrm{d}x + \frac{\partial v}{\partial x}\mathrm{d}x\mathrm{d}t\right)$$

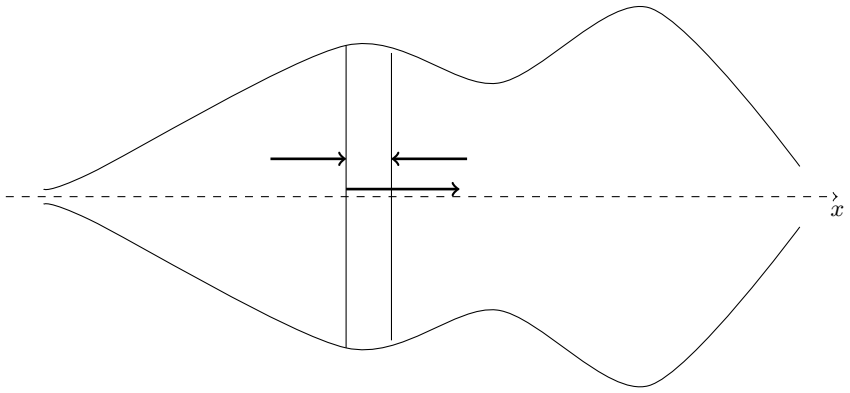


Figure 5.8: Momentum

which leads to the equation

$$\frac{\partial \rho A}{\partial x} v + \frac{\partial \rho A}{\partial t} + \rho A \frac{\partial v}{\partial x} = 0$$

Momentum: The momentum conservation principle says that $F = m\frac{\partial v}{\partial t}$. Here, the only force is the pressure force:

$$P(x,t)A(x,t) - P(x+\mathrm{d}x,t)A(x+\mathrm{d}x,t) = -\frac{\partial PA}{\partial x}\mathrm{d}x$$

while the momentum derivative is given by

$$m\frac{\partial v}{\partial t} = \rho A \mathrm{d}x \frac{\partial v}{\partial t}.$$

This leads to

$$\frac{\partial PA}{\partial x} + \rho A \frac{\partial v}{\partial t} = 0$$

Physical equations: We have thus

$$\begin{cases} \rho = \rho_0 + \rho_e & \text{and} \quad P = f(\rho_0) + \kappa^2 \rho_e \\ \frac{\partial \rho A}{\partial x} v + \frac{\partial \rho A}{\partial t} + \rho A \frac{\partial v}{\partial x} = 0 & \text{and} \quad \frac{\partial PA}{\partial x} + \rho A \frac{\partial v}{\partial t} = 0 \end{cases}$$

A classical approximation consists in neglecting some spatial deformation if front of temporal ones: $\frac{\partial \rho A}{\partial x} v \ll \frac{\partial \rho A}{\partial t}$.

$$\begin{cases} \rho = \rho_0 + \rho_e & \text{and} \quad P = f(\rho_0) + \kappa^2 \rho_e \\ \frac{\partial \rho A}{\partial t} + \rho A \frac{\partial v}{\partial x} = 0 & \text{and} \quad \frac{\partial PA}{\partial x} + \rho A \frac{\partial v}{\partial t} = 0 \end{cases}$$

Rewriting this system in ρ_e yields

$$\begin{cases} \rho_0 \frac{\partial A}{\partial t} + \frac{\partial \rho_e A}{\partial t} + \rho_0 A \frac{\partial v}{\partial x} + \rho_e A \frac{\partial v}{\partial x} = 0 \\ f(\rho_0) \frac{\partial A}{\partial x} + \kappa^2 \frac{\partial \rho_e A}{\partial x} + \rho_0 A \frac{\partial v}{\partial x} + \rho_e A \frac{\partial v}{\partial t} = 0 \end{cases}$$

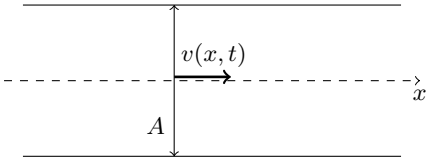


Figure 5.9: Cylinder model

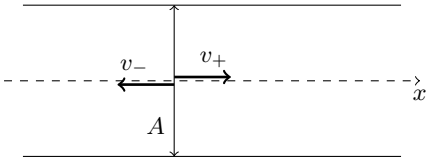


Figure 5.10: Two waves

Now neglecting ρ_e in front of ρ_0 yields

$$\begin{cases} \rho_0 \frac{\partial A}{\partial t} + \frac{\partial \rho_e A}{\partial t} + \rho_0 A \frac{\partial v}{\partial x} = 0 \\ f(\rho_0) \frac{\partial A}{\partial x} + \kappa^2 \frac{\partial \rho_e A}{\partial x} + \rho_0 A \frac{\partial v}{\partial t} = 0 \end{cases}$$

which is a complex model for which no explicit solution exists.

Cylinder model: in order to be able to compute explicit solution, we will assume that A is constant along the tube spatially and temporally. In that case, the equation system becomes

$$\begin{cases} A \frac{\partial \rho_e}{\partial t} + \rho_0 A \frac{\partial v}{\partial x} = 0 \\ \kappa^2 A \frac{\partial \rho_e}{\partial x} + \rho_0 A \frac{\partial v}{\partial t} = 0 \end{cases}$$

which implies the wave equation:

$$\frac{\partial^2 v}{\partial x^2} - \frac{1}{\kappa^2} \frac{\partial^2 v}{\partial t^2} = 0.$$

Note that the application which assign a solution to an initial condition is Linear (time) Translation Invariant...

Resolution: One computes first the Fourier transform of the system along the time:

$$\frac{\partial^2 \widehat{v}}{\partial x^2}(x,\omega) + \frac{1}{\kappa^2} \omega^2 \widehat{v}(x,\omega) = 0$$

It leads to a second order linear differential equation of solution:

$$\begin{aligned} \widehat{v}(x,\omega) &= \widehat{v^+}(\omega) e^{-i\omega x/\kappa} - \widehat{v^-}(\omega) e^{+i\omega x/\kappa} \\ v(x,t) &= \int_{\mathbb{R}} \widehat{v^+}(\omega) e^{i\omega(t-x/\kappa)} \mathrm{d}\omega - \int_{\mathbb{R}} \widehat{v^-}(\omega) e^{i\omega(t+x/\kappa)} \mathrm{d}\omega \\ &= v^+(t - x/\kappa) - v^-(t + x/\kappa) \end{aligned}$$

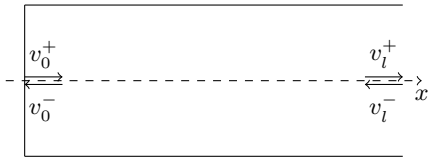


Figure 5.11: Boundary conditions

The solution can be obtained as the sum of two waves, a forward wave v^+ and a backward wave v^-, both going at speed κ.

For the air density (and the pressure), we use

$$v = v^+(t - x/\kappa) - v^-(t + x/\kappa)$$
$$\frac{\partial \rho_e}{\partial t} + \rho_0 \frac{\partial v}{\partial x} = 0$$

which yields

$$\frac{\partial \rho_e}{\partial t}(x,t) = \frac{\rho_0}{\kappa}\left(\frac{\partial v^+}{\partial t}(t - x/\kappa) + \frac{\partial v^-}{\partial t}(t + x/\kappa)\right)$$
$$\rho_e(x,t) = \frac{\rho_0}{\kappa}\left(v^+(t - x/\kappa) + v^-(t + x/\kappa)\right)$$

The pressure is thus given by

$$P(x,t) - f(\rho_0) = \rho_0\left(v^+(t - x/\kappa) + v^-(t + x/\kappa)\right)$$

Continuous physical quantities: before we consider boundary condition, we may notice that two physical quantities should remain continuous in the system:

- the air flow:

$$\rho_0 A v(x,t) = \rho_0 A v^+(t - x/\kappa) - \rho_0 A v^-(t + x/\kappa)$$

- the pressure:

$$P(x,t) - f(\rho_0) = \rho_0\left(v^+(t - x/\kappa) + v^-(t + x/\kappa)\right)$$

Boundary conditions: We focus here on two very simple boundary conditions: open end and close end. In an open end, the pressure remains constant, equal to the ambient pressure so that

$$v^+(t - x/\kappa) = -v^-(t + x/\kappa)$$
$$\widehat{v^+}(\omega)e^{i\omega x/\kappa} = -\widehat{v^-}(\omega)e^{-i\omega x/\kappa}$$

In a closed end, no air flows across the boundary so that

$$v^+(t - x/\kappa) = v^-(t + x/\kappa)$$
$$\widehat{v^+}(\omega)e^{i\omega x/\kappa} = \widehat{v^-}(\omega)e^{-i\omega x/\kappa}$$

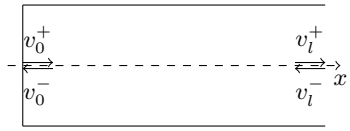


Figure 5.12: Tube as a system

More general boundary conditions can be obtained by assuming the existence of a complex impedance Z_x such that

$$\widehat{v^+}(\omega)e^{i\omega x/\kappa} = Z_x(\omega)\widehat{v^-}(\omega)e^{-i\omega x/\kappa}$$

.

5.2 Modeling and Estimating

5.2.1 System modeling

Tube as a system: Our physical modeling gives a relationship between $v_x^+(t) = v^+(t - x/\kappa)$ and $v_x^-(t) = v^+(t + x/\kappa)$ for different position. In term of $(v_0^+, v_0^-)^t$ and $(v_l^+, v_l^-)^t$, one has

$$\begin{pmatrix} v_l^+(t) \\ v_l^-(t) \end{pmatrix} = \begin{pmatrix} v^+(t - l/\kappa) \\ v^-(t + l/\kappa) \end{pmatrix} = \begin{pmatrix} v_0^+(t - l/\kappa) \\ v_0^-(t + l/\kappa) \end{pmatrix}$$

or equivalently in the Fourier domain

$$\begin{pmatrix} \widehat{v_l^+}(\omega) \\ \widehat{v_l^-}(\omega) \end{pmatrix} = \begin{pmatrix} e^{i\omega l/\kappa} & 0 \\ 0 & e^{-i\omega l/\kappa} \end{pmatrix} \begin{pmatrix} \widehat{v_0^+}(\omega) \\ \widehat{v_0^-}(\omega) \end{pmatrix}$$

We would rather look at this system the other way:

$$\begin{pmatrix} \widehat{v_0^+}(\omega) \\ \widehat{v_0^-}(\omega) \end{pmatrix} = \begin{pmatrix} e^{-i\omega l/\kappa} & 0 \\ 0 & e^{i\omega l/\kappa} \end{pmatrix} \begin{pmatrix} \widehat{v_l^+}(\omega) \\ \widehat{v_l^-}(\omega) \end{pmatrix}$$

If we want to obtain a system in term of input and output velocity, one should specify some boundary conditions.

Simple open end model: We assume that the mouth can be modeled by an open end: $v^+(t - l/\kappa) = -v^-(t + l/\kappa) = v(l, t)/2$ or equivalently if the Fourier domain $\widehat{v_l^+}(\omega) = -\widehat{v_l^-}(\omega) = \frac{1}{2}\widehat{v_l}(\omega)$. A forced velocity v_0 at 0 is thus related to v_l by:

$$\begin{aligned}
\widehat{v_0} &= \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \widehat{u_0^+} \\ \widehat{u_0^-} \end{pmatrix} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} e^{-i\omega l/\kappa} & 0 \\ 0 & e^{i\omega l/\kappa} \end{pmatrix} \begin{pmatrix} \widehat{v_l^+}(\omega) \\ \widehat{v_l^-}(\omega) \end{pmatrix} \\
&= \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} e^{-i\omega l/\kappa} & 0 \\ 0 & e^{i\omega l/\kappa} \end{pmatrix} \begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix} \widehat{v_l} \\
&= \frac{1 + e^{2i\omega l/\kappa}}{2e^{i\omega l/\kappa}} \widehat{v_l} \\
\widehat{v_0} &= \cos(\omega l/\kappa)\widehat{v_l}
\end{aligned}$$

This corresponds to a transfer function:

$$\frac{\widehat{v_l}}{\widehat{v_0}} = \frac{2e^{i\omega l/\kappa}}{1+e^{2i\omega l/\kappa}} = \frac{1}{\cos(\omega l/\kappa)}$$

in which resonance occurs for $\omega_k = \pi\kappa/l(1/2+k)$!

Attenuation: In practice, attenuation of the waves by a factor $0 \leq \alpha < 1$ when traveling from one side to the other so that the transfer matrix becomes

$$\begin{pmatrix} \widehat{v_0^+}(\omega) \\ \widehat{v_0^-}(\omega) \end{pmatrix} = \begin{pmatrix} \frac{1}{\alpha}e^{-i\omega l/\kappa} & 0 \\ 0 & \alpha e^{i\omega l/\kappa} \end{pmatrix} \begin{pmatrix} \widehat{v_l^+}(\omega) \\ \widehat{v_l^-}(\omega) \end{pmatrix}$$

Thus

$$\widehat{v_0} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{\alpha}e^{-i\omega l/\kappa} & 0 \\ 0 & \alpha e^{i\omega l/\kappa} \end{pmatrix} \begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix} \widehat{v_l}$$

$$\widehat{v_0} = \frac{-\alpha + \frac{1}{\alpha}e^{i2\omega l/\kappa}}{2e^{-i\omega l/\kappa}} \widehat{v_l}\widehat{v_0} \qquad = \left[(\alpha+1/\alpha)\cos(\omega l/\kappa) + i(\alpha-1/\alpha)\sin(\omega l/\kappa)\right]\widehat{v_l}$$

so that the transfer function is

$$\frac{\widehat{v_l}}{\widehat{v_0}} \frac{2e^{-i\omega l/\kappa}}{-\alpha + \frac{1}{\alpha}e^{-i2\omega l/\kappa}}$$

$$= \frac{2}{(\alpha+1/\alpha)\cos(\omega l/\kappa) + i(\alpha-1/\alpha)\sin(\omega l/\kappa)}$$

$$\left|\frac{\widehat{v_l}}{\widehat{v_0}}\right|^2 = \frac{1}{\frac{1/\alpha-\alpha}{4} + \cos^2(\omega l/\kappa)} \leq \frac{4}{1/\alpha - \alpha} < +\infty \quad \text{if } \alpha < 1.$$

Note that the transfer function can be written as

$$\frac{e^{-i\omega l/\kappa}}{P(e^{-i2\omega l/\kappa})}$$

with P a polynomial of degree 1. This is thus compatible with a discretization with time step $l/(N\kappa)$ with $N \in \mathbb{N}$. Indeed, assume $\widehat{v_0}$ is compactly supported in $(-\pi N\kappa/l, \pi N\kappa/l)$ then so is $\widehat{v_l}$. This implies that both v_0 and v_l can be recovered from their sampled version

$$v_{0,l/(N\kappa)} = l/(N\kappa) \sum_{n\in\mathbb{Z}} v_0(nl/(N\kappa))\delta_{nl/(N\kappa)}$$

$$v_{l,l/(N\kappa)} = l/(N\kappa) \sum_{n\in\mathbb{Z}} v_l(nl/(N\kappa))\delta_{nl/(N\kappa)}$$

which are related by

$$\mathcal{F}v_{l,l/(N\kappa)} = \frac{e^{-i\omega Kl/\kappa}}{P(e^{-i\omega 2Nl/(N\kappa)})} \mathcal{F}v_{0,l/(N\kappa)}.$$

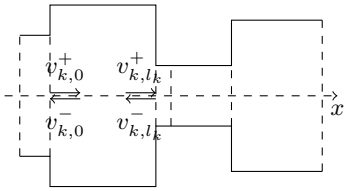


Figure 5.13: Sequences of tubes

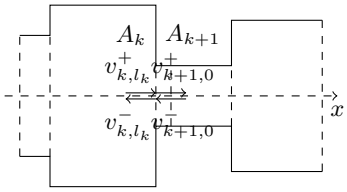


Figure 5.14: Interface

Sequence of tubes: we will model our vocal tract by a sequence of tubes of constant width. We denote by A_k the surface of the kth tube lying between $L_{k-1} = \sum_{k' \leq (k-1)} l_{k'}$ and $L_k = \sum_{k' \leq k} l_{k'}$. Its velocity is parameterized by : for $x \in (L_{k-1}, L_k)$

$$v(x,t) = v_k(x - L_{k-1}, t) = v_k^+ \left(t - \frac{x - L_{k-1}}{\kappa} \right) - v_k^- \left(t + \frac{x - L_{k-1}}{\kappa} \right)$$

and its pressure by

$$P(x,t) - f(\rho_0) = \rho_0 \left(v_k^+ \left(t - \frac{x - L_{k-1}}{\kappa} \right) + v^- \left(t + \frac{x - L_{k-1}}{\kappa} \right) \right)$$

Inside the kth tube,

$$\begin{pmatrix} \widehat{v^+_{k,0}}(\omega) \\ \widehat{v^-_{k,0}}(\omega) \end{pmatrix} = e^{i\omega l_k/\kappa} \begin{pmatrix} \frac{1}{\alpha_k} e^{-i\omega 2 l_k/\kappa} & 0 \\ 0 & \alpha_k \end{pmatrix} \begin{pmatrix} \widehat{v^+_{k,l_k}}(\omega) \\ \widehat{v^-_{k,l_k}}(\omega) \end{pmatrix}$$

Looking at the continuity of the air flow and of the pressure, we deduce

$$\begin{aligned} A_k \rho_0 \left(\widehat{v^+_{k,l_k}} - \widehat{v^-_{k,l_k}} \right) &= A_{k+1} \rho_0 \left(\widehat{v^+_{k+1,0}} - \widehat{v^-_{k+1,0}} \right) \\ \rho_0 \left(\widehat{v^+_{k,l_k}} + \widehat{v^-_{k,l_k}} \right) &= \rho_0 \left(\widehat{v^+_{k+1,0}} + \widehat{v^-_{k+1,0}} \right) \end{aligned}$$

Solving this system in v^+_{k,l_k} and v^-_{k,l_k} yields

$$\begin{aligned} \widehat{v^+_{k,l_k}} &= \left(\frac{A_k + A_{k+1}}{2A_k} \right) \widehat{v^+_{k+1,0}} + \frac{A_k - A_{k+1}}{2A_k} \widehat{v^-_{k+1,0}} \\ \widehat{v^-_{k,l_k}} &= \frac{A_k - A_{k+1}}{2A_k} \widehat{v^+_{k+1,0}} + \left(\frac{A_k + A_{k+1}}{2A_k} \right) \widehat{v^-_{k+1,0}} \end{aligned}$$

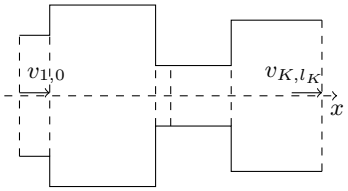


Figure 5.15: Sequence of tubes

Denoting $-1 \leq r_k = \frac{A_k - A_{k+1}}{A_k + A_{k+1}} \leq 1$ the reflection coefficient, one has thus

$$\begin{pmatrix} \widehat{v^+_{k,l_k}} \\ \widehat{v^-_{k,l_k}} \end{pmatrix} = \frac{1}{1+r_k} \begin{pmatrix} 1 & r_k \\ r_k & 1 \end{pmatrix} \begin{pmatrix} \widehat{v^+_{k+1,0}} \\ \widehat{v^-_{k+1,0}} \end{pmatrix}$$

One has thus

$$\begin{pmatrix} \widehat{v^+_{k,l_k}} \\ \widehat{v^-_{k,l_k}} \end{pmatrix} = \frac{1}{1+r_k} \begin{pmatrix} 1 & r_k \\ r_k & 1 \end{pmatrix} \begin{pmatrix} \widehat{v^+_{k+1,0}} \\ \widehat{v^-_{k+1,0}} \end{pmatrix}$$

$$\begin{pmatrix} \widehat{v^+_{k,0}} \\ \widehat{v^-_{k,0}} \end{pmatrix} = \frac{e^{i\omega l_k/\kappa}}{1+r_k} \begin{pmatrix} \frac{1}{\alpha_k} e^{-i\omega 2l_k/\kappa} & 0 \\ 0 & \alpha_k \end{pmatrix} \begin{pmatrix} 1 & r_k \\ r_k & 1 \end{pmatrix} \begin{pmatrix} \widehat{v^+_{k+1,0}} \\ \widehat{v^-_{k+1,0}} \end{pmatrix}$$

$$\begin{pmatrix} \widehat{v^+_{k,0}} \\ \widehat{v^-_{k,0}} \end{pmatrix} = \frac{e^{i\omega l_k/\kappa}}{1+r_k} \begin{pmatrix} \frac{1}{\alpha_k} e^{-i\omega 2l_k/\kappa} & \frac{r_k}{\alpha_k} e^{-i\omega 2l_k/\kappa} \\ \alpha_k r_k & \alpha_k \end{pmatrix} \begin{pmatrix} \widehat{v^+_{k+1,0}} \\ \widehat{v^-_{k+1,0}} \end{pmatrix}$$

Cascading this structure yields:

$$\begin{pmatrix} \widehat{v^+_{1,0}} \\ \widehat{v^-_{1,0}} \end{pmatrix} = \frac{e^{i\omega L_{K-1}/\kappa}}{\prod_{k'=1}^{K-1}(1+r_{k'})} \prod_{k'=1}^{K-1} \begin{pmatrix} \frac{1}{\alpha_{k'}} e^{i\omega 2l_{k'}/\kappa} & \frac{r_{k'}}{\alpha_{k'}} e^{i\omega 2l_{k'}/\kappa} \\ \alpha_{k'} r_{k'} & \alpha_{k'} \end{pmatrix} \begin{pmatrix} \widehat{v^+_{K-1,0}} \\ \widehat{v^-_{K-1,0}} \end{pmatrix}$$

As

$$\begin{pmatrix} \widehat{v^+_{K,0}} \\ \widehat{v^-_{K,0}} \end{pmatrix} = e^{i\omega l_K/\kappa} \begin{pmatrix} \frac{1}{\alpha_K} e^{-i\omega 2l_K/\kappa} & 0 \\ 0 & \alpha_K \end{pmatrix} \begin{pmatrix} \widehat{v^+_{K,l_K}} \\ \widehat{v^-_{K,l_K}} \end{pmatrix}$$

One has

$$\begin{pmatrix} \widehat{v^+_{1,0}} \\ \widehat{v^-_{1,0}} \end{pmatrix} = \frac{e^{i\omega L_K/\kappa}}{\prod_{k'=1}^{K-1}(1+r_{k'})} \prod_{k'=1}^{K-1} \begin{pmatrix} \frac{1}{\alpha_{k'}} e^{-i\omega 2l_{k'}/\kappa} & \frac{r_{k'}}{\alpha_{k'}} \\ \alpha_{k'} r_{k'} e^{-i\omega 2l_{k'}/\kappa} & \alpha_{k'} \end{pmatrix} \begin{pmatrix} \widehat{v^+_{K,l_K}} \\ \widehat{v^-_{K,l_K}} \end{pmatrix}$$

where we have set $r_K = 0$. Finally using the open end boundary condition at the end

$$\frac{\widehat{v_{1,0}}}{\widehat{v_{K,l_K}}} = \begin{pmatrix} 1 & -1 \end{pmatrix} \frac{e^{i\omega L_K/\kappa}}{\prod_{k'=1}^{K}(1+r_{k'})} \prod_{k'=1}^{K} \begin{pmatrix} \frac{1}{\alpha_{k'}} e^{-i\omega 2l_{k'}/\kappa} & \frac{r_{k'}}{\alpha_{k'}} e^{-i\omega 2l_{k'}/\kappa} \\ \alpha_{k'} r_{k'} & \alpha_{k'} \end{pmatrix} \begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix}$$

The resulting transfer function is thus

$$\frac{e^{-i\omega L_K/\kappa} / \prod_{k'=1}^{K}(1+r_{k'})}{\begin{pmatrix} 1 & 1 \end{pmatrix} \prod_{k'=1}^{K} \begin{pmatrix} \frac{1}{\alpha_{k'}} e^{-i\omega 2l_{k'}/\kappa} & \frac{r_{k'}}{\alpha_{k'}} e^{-i\omega 2l_{k'}/\kappa} \\ \alpha_{k'} r_{k'} & \alpha_{k'} \end{pmatrix} \begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix}}$$

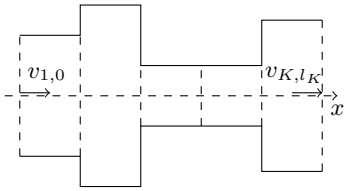


Figure 5.16: Constant length tubes

Extending to complex impedance at the end:

$$\frac{e^{-i\omega L_K/\kappa}/\prod_{k'=1}^{K}(1+r_{k'})}{\begin{pmatrix}1 & 1\end{pmatrix}\prod_{k'=1}^{K}\begin{pmatrix}\frac{1}{\alpha_{k'}}e^{-i\omega 2l_{k'}/\kappa} & \frac{r_{k'}}{\alpha_{k'}}e^{-i\omega 2l_{k'}/\kappa}\\ \alpha_{k'}r_{k'} & \alpha_{k'}\end{pmatrix}\begin{pmatrix}Z_{K,l_K}(\omega)/(Z_{K,l_K}(\omega)-1)\\ 1/(Z_{K,l_K}(\omega)-1)\end{pmatrix}}$$

Constant length tubes: A very interesting case is the one of constant length tubes $l_k = l$:

$$\begin{aligned}&\frac{e^{-i\omega Kl/\kappa}/\prod_{k'=1}^{K}(1+r_{k'})}{\begin{pmatrix}1 & -1\end{pmatrix}\prod_{k'=1}^{K}\begin{pmatrix}\frac{1}{\alpha_{k'}}e^{-i\omega 2l/\kappa} & \frac{r_{k'}}{\alpha_{k'}}e^{-i\omega 2l/\kappa}\\ \alpha_{k'}r_{k'} & \alpha_{k'}\end{pmatrix}\begin{pmatrix}1/2\\ -1/2\end{pmatrix}}\\ &= \frac{e^{-i\omega Kl/\kappa}}{P_K(e^{-i\omega Zl/\kappa})}\end{aligned}$$

where P_K is a polynomial of degree at most K.

Time discretization: By construction

$$\frac{e^{-i\omega Kl/\kappa}}{P_K(e^{-i\omega 2l/\kappa})}$$

is a $2\pi\kappa/l$ periodic function which is thus compatible with a discrete signal with sample $l/(N\kappa)$ with $N \in \mathbb{N}$. Such a compatibility is ensured in the complex impedance case if $Z_{K,l}(\omega) = W_{K,l}(e^{-i\omega l/(N\kappa)})$ (at least approximately for $|\omega| \leq \pi N\kappa/l$). Assume $\widehat{v}_{1,0}$ is compactly supported in $(-\pi N\kappa/l, \pi N\kappa/l)$ then so is $\widehat{v}_{K,l}$. Both $v_{1,0}$ and $v_{K,l}$ can be recovered from their sampled version

$$\begin{aligned}v_{1,0,l/(N\kappa)} &= l/(N\kappa)\sum_{n\in\mathbb{Z}} v_{1,0}(nl/(N\kappa))\delta_{nl/(N\kappa)}\\ v_{K,l,l/(N\kappa)} &= l/(N\kappa)\sum_{n\in\mathbb{Z}} v_{K,l}(nl/(N\kappa))\delta_{nl/(N\kappa)}\end{aligned}$$

which are related by

$$\mathcal{F}v_{K,l_K,l/(N\kappa)} = \frac{e^{-i\omega Kl/\kappa}}{P_K(e^{-i\omega 2Nl/(N\kappa)})}\mathcal{F}v_{1,0,l/(N\kappa)}$$

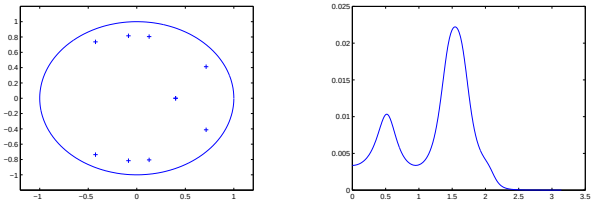


Figure 5.17: Typical AR filter with poles and squared modulus

If we denote $\mathrm{v}_{1,0}[n] = v_{1,0}(nl/(N\kappa))$ and $\mathrm{v}_{K,l}[n] = v_{1,0}(nl/(N\kappa))$, one deduces

$$\frac{\widehat{\mathrm{v}_{K,l}}(e^{-i\omega})}{\widehat{\mathrm{v}_{1,0}}(e^{-i\omega})} = \frac{e^{-i\omega KN}}{P_K(e^{-i\omega 2N})}$$

which, up to a fixed delay of KN, corresponds to an AR filter of size $2KN$ with value separated by $2N$. Such a filter can be implemented by an AR filter with K parameters!

Finally, in the case of complex impedance Z for the boundary condition:

$$\frac{\widehat{\mathrm{v}_{K,l}}(e^{-i\omega})}{\widehat{\mathrm{v}_{1,0}}(e^{-i\omega})} = \frac{e^{-i\omega KN} Z_{K,l}(e^{i\omega}}{P_K^1(e^{-i\omega 2N}) + P_K^2(e^{-i\omega 2N}) Z_{K,l}(e^{i\omega})}$$

which is an AR filter up to a know FIR if $Z_{K,l}$ is a polynomial in $e^{i\omega}$ (or better in $e^{i\omega 2N}$)!

5.2.2 Voiced and unvoiced case modeling and estimation

Unvoiced case modeling: The vocals folds are not oscillating: the excitation $v_{1,0}$ is a random noise whose power spectrum is supported in $(-\pi N\kappa/l, \pi N\kappa/l)$. Its discretization $\mathrm{v}_{1,0}$ with step $l/(N\kappa)$ is thus a discrete random noise (a zero mean WSSP) whose power spectrum density $\widehat{R_{\mathrm{v}_{1,0}}}(e^{-i\omega})$ is supported in $(-\pi, \pi)$. By construction, $\mathrm{v}_{K,l}$ is thus a WSSP whose power spectrum density is given by

$$\widehat{R_{\mathrm{v}_{K,l}}}(e^{-i\omega}) = \frac{1}{|P_K(e^{-i2N\omega})|^2} \widehat{R_{\mathrm{v}_{1,0}}}(e^{-i\omega})$$

If we do a further approximation, the one that $\mathrm{v}_{1,0}$ is a noise whose spectrum is the inverse of a polynomial

$$\widehat{R_{\mathrm{v}_{1,0}}}(e^{-i\omega}) = \frac{1}{|Q(e^{-i2N\omega})|^2} \Rightarrow \widehat{R_{\mathrm{v}_{K,l}}}(e^{-i\omega}) = \frac{1}{|P_K(e^{-i2N\omega})Q(e^{-i2N\omega})|^2}$$

then $\mathrm{v}_{K,l}$ is an AR process!

Unvoiced case synthesis: As it is sufficient to reproduce a random signal having the same frequential behavior, it suffices to generate a white noise of the right variance and to apply the AR filter to this noise.

Unvoiced case estimation: We are exactly in the setting of the estimation of the parameters of an AR so that the Least Square method of the previous chapter can be used.

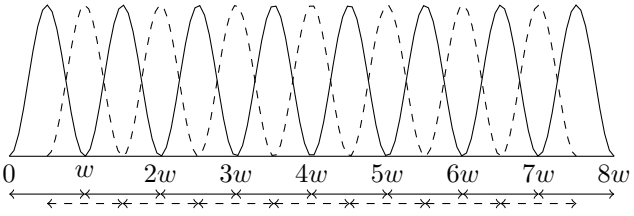


Figure 5.18: Windowing principle

Voiced case modeling: Now the vocal folds are oscillating: the excitation $v_{1,0}$ is a periodic signal of period $P_v \gg l/(N\kappa)$ and band-limited in $(-\pi N\kappa/l, \pi N\kappa/l)$ If we assume that $P_v = Pl/(N\kappa)$ with $P \in \mathbb{N}$, this periodicity corresponds to a periodicity of $P \gg 1$ samples compatible with the time discretization of $l/(N\kappa)$ The Fourier transform of $\mathrm{v}_{1,0}$ is thus discrete of step $2\pi/P$. $\mathrm{v}_{K,l}$ has thus the same properties and its Fourier transform satisfies

$$\widehat{\mathrm{v}_{K,l}}(e^{-in2\pi/P}) = \frac{e^{-inKN2\pi/P}}{P_K(e^{-in4N\pi/P})}\widehat{\mathrm{v}_{1,0}}(e^{-in2\pi/P})$$

Again if we assume that $\widehat{\mathrm{v}_{1,0}}(e^{-in2\pi/P})$ is the inverse of a polynomial:

$$\widehat{\mathrm{v}_{1,0}}(e^{-in2\pi/P}) = \frac{1}{Q(e^{-in4\pi/P})} \Rightarrow \widehat{\mathrm{v}_{K,l}}(e^{-in2\pi/P}) = \frac{e^{-inKN2\pi/P}}{P_K(e^{-in4N\pi/P})Q(e^{-in4N\pi/P})}$$

$\mathrm{v}_{K,L}$ is, up to a delay, approximately the convolution of an AR filter with a Dirac train of period P!

Voiced case synthesis: It suffices to generate a Dirac train of period P and correct amplitude and to apply the AR filter to this train.

Voiced case estimation: Both the period and the AR parameters should be estimated. Several methods have been proposed to estimate the period, the easiest one is based on the computation of

$$\Delta(p) = \frac{1}{M}\sum_n |\mathrm{v}_{K,l}[n] - \mathrm{v}_{K,l}[n-p]|.$$

If the minimum is small and attained at P, the sound is voiced with a period P otherwise it is unvoiced... AR filter parameter estimation by the same least square method used for random process! The theoretical justification is complex. Heuristically, in

$$P[1]\mathrm{v}_{K,l}[n] = -\sum k = 1^L P[k]\mathrm{v}_{K,l}[n-1] + sum_m \delta_{mP}$$

the part $-\sum k = 1^L P[k]\mathrm{v}_{K,l}[n-1]$ is much bigger than 1 due to the resonance phenomena and thus the least square fit is only mildly perturbed by neglecting this part!

Windowing: the stationarity assumption does not hold in practice but it holds approximately locally. Using the same windowing technique than the one used if the Short Time Fourier Transform, one can split the signal in overlapping piece in which the signal is multiplied by a suitable window. The LPC model is used on each piece, called a frame, and the reconstruction is obtained by combining those reconstruction using the same window.

5.3 Linear Predictive Coding

This model can be used to compress a speech signal. Instead of sending the values of v, one can send the parameters of the model (the AR coefficients, the power and the pitch) and use those parameters to synthesize an approximated version of the speech. This compression algorithm principle has been introduced in the late 60's and popularized in the 70's. It is still used for speech coding for instance in VoIP products.

The most classical version is called LPC10 and used a 10 order model with windows of size 180. It transmits 2400 b/s for a signal sampled at 8 kHz and produces a reconstruction of good quality. An important issue is the quantization of the AR parameters. For stability reason, it turns out that it is much better to quantize not the coefficients themselves but a reparameterization based on the reflectivity coefficients seen in the tube model.

The residual, the error between the reconstruction and the original signal, can be further reduced by selecting the excitation signal in a dictionary instead of either a Dirac train or a purely random noise. This technique called CELP (Code Excited Linear Predicting) is the one used in speech codec today. If one wants to compress other types of sound, the model is not accurate anymore and one has to resort to the information theory technique presented in the next chapter.

Part III

Image compression: Information Theory and Transform Coding

In this last part, we will study how to *compress* images. We will consider two forms of compression: the lossless coding in which no information loss occurs and the lossy ones in which one trades more compression with a loss of quality.

Chapter 6

Lossless coding

6.1 Coding

Symbols and words: More generally, we consider objects that can be described by words, i.e. finite sequences of symbols. We let $\mathcal{A} = \{a_k\}_{1 \leq k \leq K}$ be a dictionary of K symbols and denote w a word of size $|w|$, i.e. a finite sequence $w^1 w^2 \dots w^{|w|}$ with $w^i \in \mathcal{A}$.

Set of words: Objects of interest belong to subsets $\mathcal{W}$ of $|\mathcal{W}|$ words:

- Words of 1 symbol: $\mathcal{W} = \mathcal{A}$ with $|\mathcal{W}| = |\mathcal{A}|$.
- Words of n symbols: $\mathcal{W} = \mathcal{A}^n$ with $|\mathcal{W}| = |\mathcal{A}|^n$.
- Words of less than n symbols: $\mathcal{W} = \cup_{k=1}^{n} \mathcal{A}^n = \mathcal{A}^{*n}$ with $|\mathcal{W}| = \frac{|\mathcal{A}|^{n+1} - |\mathcal{A}|}{|\mathcal{A}| - 1}$.
- Words of arbitrary size: $\mathcal{W} = \cup_{k=1}^{+\infty} \mathcal{A}^n = \mathcal{A}^*$ with $|\mathcal{W}| = +\infty$

Code: a code C assign a finite length binary sequence to any word of the subset $\mathcal{W}$:

$$\begin{aligned} \mathcal{W} &\to \{0,1\}^* \\ w &\mapsto C(w) \end{aligned}$$

Lossless coding means perfect reconstruction, i.e. that C should be injective:

$$C(w_1) = C(w_2) \implies w_1 = w_2.$$

Such a code is said to be non-singular.

Fundamental list example: the simplest code is obtained by an enumeration principle

- Chose l such that $2^l \geq |\mathcal{W}|$.
- List all words of $\mathcal{W}$ in an arbitrary order.

- List all binary words of $\{0,1\}^l$ in an arbitrary order.
- Assign to the ith word in the list of $\mathcal{W}$ the ith in the list of $\{0,1\}^l$

Proposition 6.1: This code is non singular.

Code length: For any code C, the length of the codeword $C(w)$ is denoted by $l(w)$. For instance, for the fundamental list example

- $\forall w \in \mathcal{W}, l(w) = l$.
- Smallest size obtained for l, $l = \lceil \log_2 |\mathcal{W}| \rceil$

Code efficiency: a good code will be one whose length is small.

Proposition 6.2 (Pigeonghole principle): If C is a code from $\{0,1\}^{*,n}$ into $\{0,1\}^*$ such that $\exists w, l(w) < |w|$ then $\exists w', l(w') > |w'|$

To compress some strings, one should expand other ones!

Shannon approach: Assume there is a (known) probability law P on the words and mesure the efficiency of a code by its average length:

$$\sum_{w \in \mathcal{W}} P(w) l(w).$$

6.2 Compression limit and entropy

Entropy as a lower bound: the entropy $H(P)$ of P is defined by

$$H(P) = -\sum_{w \in \mathcal{W}} P(w) \log_2 P(w)$$

appears as a lower bound on the efficiency of a code.

Proposition 6.3: For any non singular code C

$$\sum_{w \in \mathcal{W}} P(w) l(w) \geq H(P) - \log_2 \lceil \log_2 |\mathcal{W}| \rceil$$

Later, we will show that if one is restricted to *interesting* code,

$$\sum_{w \in \mathcal{W}} P(w) l(w) \geq H(P)$$

and that among those *interesting* code there is one such that

$$\sum_{w \in \mathcal{W}} P(w) l(w) \leq H(P) + 1$$

Entropy and complexity: Let P be any law defined on a discrete set Ω, its entropy $H(P)$ is defined by

$$H(P) = -\sum_{\omega\in\Omega} P(\omega)\log_2 P(\omega).$$

Proposition 6.4:

- $0 \leq H(P) \leq \log_2 |\mathcal{W}|$
- $H(P) = 0 \implies \exists w, P(w) = 1$
- $H(P) = \log_2 |\mathcal{W}| \implies \forall w, P(w) = 1/|\mathcal{W}|$

H is a measures of the complexity of P, which is maximal when all events are equiprobale and minimal when on is certain.

Entropy and information content: Let I be a function measuring the *information content* of $A \subset \mathcal{P}(\Omega)$.

Proposition 6.5: If we assume that

- I is a non increasing continuous function of $P(A)$
- $I(A_1 \cup A_2) = I(A_1) + I(A_2)$ as soon as A_1 and A_2 are independent

then $I(A) = -\kappa \log_2 P(A)$ with $\kappa > 0$.

Corollary 6.1: $H(P)$ is the average information content in the atomic event of Ω for the choice $\kappa = 1$.

Compression limite and entropy: For any code C on $\mathcal{W}$, we denote by $\Sigma(C) = \sum_{w\in\mathcal{W}} 2^{-l(w)} \leq |\mathcal{W}|$ and for any set $\mathcal{C}$ of codes, $\Sigma(\mathcal{C}) = \sup_{C\in\mathcal{C}} \Sigma(C)$

Theorem 6.1 (Shannon):

$$\min_{C\in\mathcal{C}} \sum_{w\in\mathcal{W}} P(w)l(w) \geq H(P) - \log_2(\Sigma(\mathcal{C})).$$

Proof: Let $\Sigma(C) = \sum_{w\in\mathcal{W}} 2^{-l(w)}$. By construction, $q(w) = 2^{-l(w)}/\Sigma(C)$ define a probability law of w.

Now, recall that the Kullback-Leibler divergence $KL(p,q) = -\sum_{w\in\mathcal{W}} p(w)\log_2 q(w)/p(w)$ satisfies $KL(p,q) \geq 0$ and thus with our choice of q

$$KL(p,q) = \sum_{w\in\mathcal{W}} p(w)l(w) + \log_2 \Sigma(C) - H(X) \geq 0$$

Proof (Kullback-Leibler positivity): By convexity of $-\log_2(x)$ and Jensen inequality

$$KL(p,q) = -\sum_{w\in\mathcal{W}} p(w)\log_2\left(q(w)/p(w)\right) = -\sum_{w,p(w)>0} p(w)\log_2\left(q(w)/p(w)\right)$$
$$\leq -\log_2\left(\sum_{w,p(w)>0} p(w)\left(q(w)/p(w)\right)\right) \geq -\log_2\left(\sum_{w,p(w)} q(w)\right) \geq 0$$

Proposition 6.6: If a code is non singular then $\Sigma(C) \leq \lceil \log_2 |\mathcal{W}| \rceil$.

Proof: Let C be a non singular code for $\mathcal{W}$ and $l_{\max}$ its maximal length, Let $N(l)$ be the number of words whose codelength is equal to l so that

$$\Sigma(C) = \sum_{w \in \mathcal{W}} 2^{-l(w)} = \sum_{l=1}^{l_{\max}} 2^{-l} N(l)$$

By construction $\sum_{l=1}^{+\infty} N(l) = |\mathcal{W}|$ and as all codes should be different $0 \leq N(l) \leq 2^l$

Let l' be the smallest integer such that $N(l') < 2^{l'}$

- if $l' < l_{\max}$ then one can construct a new non singular code C' such by replacing a code of length $> l'$ by a code of length l' so that $\Sigma(C') > \Sigma(C)$
- if $l' \geq l_{\max}$ then

$$\Sigma(C) = \sum_{l=1}^{l_{\max}} 2^{-l} N(l) \leq l_{\max}$$

We have thus for the optimal bound

$$\sum_{l=1}^{l_{\max}-1} 2^{l'} + N(l_{\max}) = |\mathcal{W}|$$

If $l' = l_{\max}$ then $1 \leq N(l_{\max}) < 2^{l_{\max}}$ thus

$$\sum_{l=1}^{l_{\max}-1} 2^{l'} \leq |\mathcal{W}| - 1 \Leftrightarrow 2^{l_{\max}} - 1 \leq |\mathcal{W}| - 1$$

which implies $l_{\max} \leq \log_2 |\mathcal{W}|$.

If $l' = l_{\max} + 1$ then $|\mathcal{W}| = 2^{l_{\max}+1} - 1$ which implies

$$l_{\max} \leq \log_2(|\mathcal{W}| + 1) - 1 \leq \log_2 |\mathcal{W}|.$$

Corollary 6.2: If $\mathcal{C}$ is a set of non singular code

$$\min_{C \in \mathcal{C}} \sum_{w \in \mathcal{W}} P(w) l(w) \geq H(P) - \log_2 \lceil \log_2 |\mathcal{W}| \rceil$$

A natural question is whether those lower bound can be attained.

6.3 Typical set and typical coding

i.i.d. setting: We focus on the case where $\mathcal{W} = \mathcal{A}^n$ with $P_n(w) = \prod_{i=1} P(w_i)$.

Proposition 6.7: $H(P_n) = nP(H)$

Proposition 6.8: Thanks to the Large Number Law

$$\frac{1}{n} \sum_{i=1}^{n} -log_2 P(w_i) \to H(P)$$

ϵ-typical set: the ϵ-typical set

$$\mathcal{A}_\epsilon^n = \{w \in \mathcal{A}^n, |\frac{1}{n}\sum_{i=1}^{n} -\log_2 P(w_i) - H(P)| \leq \epsilon\}$$

satisfies $\mathbb{P}(\mathcal{A}_\epsilon^n) \to 1$.

For any $\epsilon > 0$, it exists n_0 such that for $n \geq n_0$, with probability larger than $1-\epsilon$, a random word of size n belongs to this set...

Asymptotic Equipartition Principle: more precisely

Theorem 6.2: $\forall \epsilon > 0, \exists n_0$ such that $\forall n \geq n_0$,

- $\forall w \in \mathcal{A}_\epsilon^n, 2^{-n(H(X)+\epsilon)} \leq P_n(w) \leq 2^{-n(H(P)-\epsilon)}$
- $\mathbb{P}(\mathcal{A}_\epsilon^n) \geq 1-\epsilon$
- $(1-\epsilon)2^{n(H(P)-\epsilon)} \leq |\mathcal{A}_\epsilon^n| \leq 2^{n(H(P)+\epsilon)}$

With probability close to 1, a random sequence w belongs to a set of size of order $2^{nH(P)}$ where all sequences have probability of order $2^{-nH(P)}$. In other words, $\mathcal{A}^n$ behaves similarly to a set of size $2^{nH(P)}$ of equiprobable words!

Basic list code:

- List all words of $\mathcal{A}^n$ and list all binary words of $\{0,1\}^{\lceil n \log |\mathcal{A}| \rceil}$.
- Assign to the ith word in the list of $\mathcal{A}^n$ the ith in the list of $\{0,1\}^{\lceil n \log |\mathcal{A}| \rceil}$
- Each word is coded by those $\lceil n \log |\mathcal{A}| \rceil$ bits and thus whatever P

$$\frac{1}{n}\sum_{w \in \mathcal{A}^n} P_n(w) l(w) = \frac{\lceil n \log |\mathcal{A}| \rceil}{n} \leq \log |\mathcal{A}| + 1$$

Typical code:

- List all words of $\mathcal{A}^n$, list all binary words of $\{0,1\}^{\lceil n \log |\mathcal{A}| \rceil}$, list of all words of $\mathcal{A}_\epsilon^n$ and list all words of $\{0,1\}^{\lceil n(H(P)+\epsilon) \rceil}$.
- For any word w, if it belongs to $\mathcal{A}_\epsilon^n$ assign it the code 0 followed by the corresponding word of $\{0,1\}^{\lceil n(H(P)+\epsilon) \rceil}$. Otherwise assign it 1 followed by the corresponding word in $\{0,1\}^{\lceil n \log |\mathcal{A}| \rceil}$
- The code is non singular and

$$\begin{aligned}\frac{1}{n}\sum_{w \in \mathcal{A}^n} P_n(w) l(w) &= \frac{1}{n}\left((1-\epsilon)(1+\lceil n(H(P)+\epsilon) \rceil) + \epsilon(1+\lceil n \log |\mathcal{A}| \rceil)\right) \\ &\leq H(P) + \epsilon(\log |\mathcal{A}| - H(P)) + \frac{1-\epsilon}{n} = H(P) + o(1)\end{aligned}$$

With this simple strategy, the average length per symbol can be arbitrarily close to the lower bound given by the average entropy.

Typical coding turns out to be complex to use as it requires the use of two very long lists and is rarely used in practice.

6.4 Extension code and prefix code

Extension code: a much more practical coding scheme would be one such that if $w = w^1 \dots w^{|w|}$ then

$$C(w) = C(w^1) \cdots C(w^{|w|}).$$

Indeed, in that case, it suffices to specify a code for the indivudual symbols to know the code for the words. Such a code is called an extension code.

Proposition 6.9: If $P_n = \prod^n P$ then

$$\frac{1}{n} \sum_{w \in \mathcal{A}^n} P(w) l(w) = \sum_{k=1}^{K} P(a_k) l(a_k)$$

Proof:

$$\begin{aligned} \sum_{w \in \mathcal{A}^n} P(w) l(w) &= \sum_{w^1 \in \mathcal{A}} \cdot \sum_{w^n \in \mathcal{A}} \left(\prod_{i=1}^{n} P(w^i) \times \sum_{j=1}^{n} l(w^j) \right) \\ &= \sum_{j=1}^{n} \sum_{w^j \in \mathcal{A}} P(w^j) l(w^j) = n \sum_{k=1}^{K} P(a_k) l(a_k) \end{aligned}$$

Uniquely decodale code: any non singular extension code is called a uniquely decodable code. Those code are easy to use in the coding part but may require to read the whole binary sequence in the decoding part.

Prefix code: A code such that for any symbol a_k there is no other symbol $a_{k'}$ such that $C(a_k)$ is the prefix (the beginning) of $C(a_{k'})$ is called a prefix code.

Proposition 6.10: A prefix code is uniquely decodable.

Furthermore, the decoding of a sequence is easy as one always knows when reading whether a symbol should be output or not.

Prefix code, tree and intervals: there is a correspondence between prefix code, subset leaves of binary trees and specific subsets of dyadic intervals.

Proposition 6.11: The following sets are in bijection:

- Prefix codes of size K: set of sets of K binary string c_k of length l_k such that no c_k is a strict prefix of c'_k
- Set of subsets of K leaves of a finite binary tree (c_k is the leaf label and l_k is the depth)
- Set of sets of K disjoint intervals $[\gamma_k 2^{-l_k}, (\gamma_k + 1)2^{-l_k})$ with $\gamma_k \in \{0, \dots 2^{l_k} - 1\}$ (c_k is the binary decomposition of γ_k prefixed by sufficiently 0 to be of length l_k)

Kraft: a fundamental result on prefix code is

Theorem 6.3: If C is a prefix code then

$$\sum_{k=1}^{K} 2^{-l(a_k)} \leq 1$$

and conversely if

$$\sum_{k=1}^{K} 2^{-l_k} \leq 1$$

then there is a prefix code C such that $|C(a_k)| = l_k$.

Proof: Let $I_k = [C(a_k)2^{-l(a_k)}, (C(a_k)+1)2^{-l(a_k)})$, as C is a prefix code those intervals are disjoint and included in $[0,1]$. Thus, denoting by $|I|$ the width of an interval I,

$$\sum_{k=1}^{K} 2^{-l(a_k)} = \sum_{k=1}^{K} |I_k| \leq 1.$$

Conversely, assume l_k are such that $\sum_{k=1}^{K} 2^{-l_k} \leq 1$, we may assume without loss of generality that $l_1 \leq l_2 \leq \cdots l_K$. Denote $L_k = \sum_{k'=1}^{k} 2^{-l_{k'}}$ and $L_0 = 0$ and let $I_k = [L_{k-1}, L_k)$. By construction, those intervals are disjoint, included in $[0,1)$ and of length 2^{-l_k}. Now

$$\gamma_k = 2^{l_k} L_{k-1} = 2^{l_k} \sum_{k'=1}^{k-1} 2^{-l_{k'}} = \sum_{k'=1}^{k-1} 2^{l(k)-l(k')} \in \mathbb{N}$$

and, as $k' \leq k$, $l(k) - l(k') \in \mathbb{N}$ so that $I_k = [\gamma_k 2^{-l_k}, (\gamma_k + 1)2^{-l_k})$. The set of intervals I_k has thus a corresponding prefix code.

Theorem 6.4 (Shannon): If C is a prefix code then

$$\frac{1}{n} \sum_{w \in \mathcal{A}^n} P_n(w) l(w) \geq H(P)$$

and it exists a prefix code C such that

$$\frac{1}{n} \sum_{w \in \mathcal{A}^n} P_n(w) l(w) \leq H(P) + 1$$

Proof: We start first by recalling that by construction

$$\frac{1}{n} \sum_{w \in \mathcal{A}^n} P_n(w) l(w) = \sum_{k=1}^{K} P(a_k) l(a_k).$$

Now, let $\mathcal{C}$ be the set of prefix codes, using notation of Theorem 6.1,

$$\Sigma(\mathcal{C}) = \sup_{C \in \mathcal{C}} \sum_{k=1}^{K} 2^{-l(a_k)} \leq 1$$

and thus

$$\inf_{C\in\mathcal{C}} \sum_{k=1}^{K} P(a_k) l(a_k) \geq H(P) - \log_2(\Sigma(\mathcal{C})) = H(P)$$

By construction, if we let $l_k = \lceil -\log_2 p(a_k) \rceil$ then

$$\sum_{k=1}^{K} 2^{-l_k} = \sum_{k=1}^{K} 2^{-\lceil -\log_2 p(a_k)\rceil} \leq \sum_{k=1}^{K} 2^{\log_2 p(a_k)} = \sum_{k=1}^{K} p(a_k) = 1$$

thus it exists a prefix code C such that $l(a_k) = l_k = \lceil -\log_2 p(a_k) \rceil \leq -\log_2 p(a_k) + 1$. Using this code, one obtains immediately

$$\begin{aligned} \sum_{k=1}^{K} p(a_k) l(a_k) &\leq \sum_{k=1}^{K} P(a_k)(-\log_2 P(a_k) + 1) \\ &\leq H(P) + 1 \end{aligned}$$

Optimality of the bound: The only code for $\mathcal{A} = \{0, 1\}$ is a code of constant length 1 but

$$H(P) = P(0)\log_2 P(0) + (1 - P(0))\log_2(1 - P(0)) \overset{P(0)\to 0}{\to} 0.$$

Theorem 6.5 (Mc-Millan): If C is a uniquely decodable code then

$$\frac{1}{n} \sum_{w\in\mathcal{A}^n} P_n(w) l(w) \geq H(P)$$

Not much loss by considering only prefix code!

The proof is a direct consequence of

Proposition 6.12: If C is a uniquely decodable code then

$$\sum_{k=1}^{K} 2^{-l(a_k)} \leq 1$$

Proof: As C is a uniquely decodable code, by definition $C(w) = C(w^1)\cdots C(w^n)$ is a non-singular code for $\mathcal{W} = \mathcal{A}^n$. We have seen that this implies

$$\sum_{w\in\mathcal{W}} 2^{-l(w)} \leq \log_2 |\mathcal{W}| = n \log_2 |\mathcal{A}|.$$

As C is an extension code

$$\sum_{w\in\mathcal{A}^n} 2^{-l(w)} = \sum_{w^1\in\mathcal{A}} \sum_{w^2\in\mathcal{A}} \cdots \sum_{w^n\in\mathcal{A}} 2^{-\sum_{i=1}^{n} l(w^i)} = \left(\sum_{k=1}^{K} 2^{-l(a_k)} \right)^n$$

This implies

$$\sum_{k=1}^{K} 2^{-l(a_k)} \leq (n \log_2 |\mathcal{A}|)^{1/n} = 2^{(\log_2 n + \log_2 \log_2 |\mathcal{A}|)/n} \overset{n\to\infty}{\to} 1$$

6.5 Practical codes

We assume first (without loss of generality) that $p(a_1) \geq \cdots \geq p(a_k)$.

Rounded code: let $l_k = \lceil -\log_2 p(a_k) \rceil$ so that $l_1 \leq \cdots \leq l_K$. We can rely on the construction of the previous section to obtain a good code. It suffices to let $L_0 = 0$ and $L_k = \sum_{k'=1}^{k} 2^{-l_{k'}}$ and define $C(a_k)$ as $2^{l_k} L_{k-1}$ prefixed by sufficiently 0 to be of length l_k.

Shannon code: Shannon proposes a similar construction relying on the cumulative probability function. Let $l_k = \lceil -\log_2 p(a_k) \rceil$ and define $R_0 = 0$ and $R_k = \sum_{k'=1}^{k} p(a_{k'})$, the cumulative probability functions evaluated at the jumps. We can then define $C(a_k)$ as $\lfloor 2^{l_k} R_{k-1} \rfloor$ prefixed by sufficiently 0 to be of length l_k (i.e. the l_k first binary digits of R_{k-1})

Proof: We will prove that the intervals $I_k = [C(a_k)2^{-l_k}, (C(a_k)+1)2^{-l_k})$ are disjoint as this suffices here to obtain the prefix property of the code. Notice that $(C(a_k)+1)2^{-l_k} \leq R_{k-1} + 2^{-l_k} \leq R_k$. Furthermore as $l_{k+1} \geq l_k$, $C(a_{k+1})2^{-l_{k+1}} \geq (C(a_k)+1)2^{-l_k}$ which concludes the proof

Proposition 6.13: For both code:

$$H(P) \leq \frac{1}{n} \sum_{w \in \mathcal{A}^n} P_n(w) l(w) < H(P) + 1$$

Assuming that the probability are sorted is not an issue in theory but may becomes a practical one if the number of symbols is large.

Shannon-Fano code: Fano proposes a variation of Shannon code that do not requires $p(a_1) \geq \cdots \geq p(a_k)$ at the price of a large code. Let $l_k = \lceil -\log_2 p(a_k) \rceil + 1$, $R_0 = 0$ and $R_k = \sum_{k'=1}^{k} p(a_{k'})$, one defines $C(a_k) = \lceil 2^{l_k} R_{k-1} \rceil$ prefixed with sufficiently 0 to be of length l_k.

Proof: We will prove that the intervals $I_k = [C(a_k)2^{-l_k}, (C(a_k)+1)2^{-l_k})$ are disjoint as this suffices here to obtain the prefix property of the code. Note that $R_k + 2^{-lk} \leq C(a_k)2^{-l_k}$ and thus m $R_k \leq C(a_k)2^{-l_k} \leq (C(a_k)+1)2^{-l_k} \leq R_k + 22^{l_k} \leq R_{k+1}$ which concludes the proof.

Proposition 6.14: For this code

$$H(P) \leq \frac{1}{n} \sum_{w \in \mathcal{A}^n} P_n(w) l(w) < H(P) + 2$$

Shannon code yields an average length smaller than $H(P) + 1$ but are not necessarily the most efficient one.

Huffman code: Huffman propose an explicit construction of an optimal code construction of an optimal prefix code C_H, i.e. a code such thay for any prefix code C:

$$\sum_{k=1}^{K} P(a_k) l(a_k) \geq \sum_{k=1}^{K} P(a_k) l_H(a_k).$$

Not that there is no unicity of such a code. For instance, the roles of 0 and 1 in the binary sequences can be exchanged.

Huffman tree: Huffman code is obtained through the recursive construction of a dyadic tree associated to the dictionary $\mathcal{A}$ of size K.

- Start from the dictionary made of all symbols
- At each step,
  - search for the two least probable symbols in the current dictionary
  - remove those two symbols from the dictionary and add a new symbol corresponding to their union
  - represent this operation by drawing two branches going from this new symbol to the two removed ones
- In K steps, the dictionary is empty and thus all the original symbols are the leaves of a dyadic tree.

The proof of the optimality is also obtained by recursion.

Average loss and grouping: Shannon proof gives the existence of a prefix code C for the dictionary $\mathcal{A}$ such that

$$H(P) \leq \frac{1}{n} \sum_{w \in \mathcal{A}^n} P_n(w) l(w) \leq H(P) + 1.$$

Assume we define the dictionary as $\mathcal{A}^n$, the very same theorem yields the existence of a prefix code for the dictionary $\mathcal{A}^n$ such that

$$H(P_n) = nH(P) \leq \sum_{w \in \mathcal{A}^n} P_n(w) l(w) \leq H(P_n) + 1 = nH(P) + 1.$$

Dividing the previous inequalities by n shows the upper bound of the coding loss goes from 1 bit per symbol in the first case to $1/n$ in the second case. This later strategy is nevertheless rarely used in practice as the cardinality of $\mathcal{A}^n$ can be very large making the code construction hardly feasible.

Block coding: If one code a word of $\mathcal{A}^n$ as a word of n/B symbols in $\mathcal{A}^B$ (assuming for sake of simplicity that B divides n) using a good prefix code C_B for $\mathcal{A}^B$

$$H(P) \leq \frac{1}{n} \sum_{w \in \mathcal{A}^n} P_n(w) l_B(w) \leq H(P) + \frac{1}{B}.$$

The larger B the closer we are from the optimal bound!

6.5.1 Non i.i.d. case, arithmetic coding and dictionary approach

Non i.i.d case: Our asymptotic analysis has been base on the assumption that $P_n(w) = \prod_{i=1}^n P(w^i)$ (i.i.d assumption) but using $\mathcal{A}^n$ as a dictionary, as in the previous section, always yields that the best uniquely decodable code is such that

$$H(P_n) \leq \sum_{w \in \mathcal{A}^n} P_n(w) l(w) \leq H(P_n) + 1.$$

Asymptotic behavior: If $\frac{1}{n}H(P_n) \to H$, the best uniquely decodable codes are such that

$$\frac{1}{n}\sum_{w\in\mathcal{A}^n} P_n(w)l(w) \to H$$

For instance, this properties holds if we assume that $(w^i)_i$ is a Markovian process.

Markovian modeling: to illustrate the importance of going beyong the i.i.d. case, we give here some typical sequence obtained by Shannon using different Markovian model of the English language:

- i.i.d modeling:
  OCRO HLO RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL
- Second orderd markov modeling on character:
  IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE
- Second order Markov modeling on words:
  THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

Better modeling yields a better comprehension because less sequences of the typical set are useless...

Practical? optimal codes: a natural question is how to construct a uniquely decodale code optimal up to $1/n$ bit as such a code exists. We already know to way of constructing such a code using Shannon coding of Kraft coding. However as explain before, the construction of such codes requires sorting the symbols and thus becomes hardly tractable when n is large and thus cannot be used in practice.

Individual coding: using $P_n(w) = \prod_{k=1}^n P(w^k|w^1,\dots,w^{k-1})$, one can try to code each letter w^k of the word w using the law $P(\cdot|w^1,\dots,w^{k-1})$. Both Shannon and Kraft strategies require to sort the letter at each step and thus are still computationaly heavy. Furthermore, the loss is up to 1 bit per symbol. Note that using the block coding strategy, with blocks of size B, reduces this loss to $1/B$ bit per symbol at a larger computational price.

Arithmetic coding: the most used strategy in this setting is to use a clever implementation of the Shannon-Fano coding scheme. The code obtained guarantees only a loss of up to $2/n$ bits with respect to the optimal one but its implementation is much more simple than the one of Shannon code or Kraft code. This clever strategy is based on a recursive construction of the interval

$$I_n(w) = \left[\sum_{w'<w} P_n(w'), \sum_{w'\leq w} P_n(w')\right)$$

used in Shannon-Fano. This construction is based on the observation that

$$I_k(w^1 \cdot w^k) = \sum_{k'=1}^{k} \prod_{k''=1}^{k'-1} P_{k''}(w^{k''}|w^1 \cdots w^{k''-1}) \sum_{a<w^{k'}} P_{k'}(a|w^1 \cdots w^{k'-1}) + \prod_{k'=1}^{k-1} P_{k'}(w^{k'}|w^1 \cdots w^{k'-1}) \left[0, P_k(w^k|w^1 \cdot w^{k-1})\right).$$

One can thus compute sequentially $I_1 \supset I_2 \supset \cdots \supset I_n$ without having to compute $P_n(w)$ for all words. Furthermore, b bits can already be output as soon as I_k belongs to a dyadic interval of size 2^{-b}. Arithmetic coding is a clever implementation of this idea that avoid precision issues and do not require to read all the data to start encoding. It is particularily well adapted to Markovian modeling in which $P(w) = \prod_{i=1}^{|w|} P(w_i|w_1 \cdots w_{i-m})$.

Adaptive approach: In practice, the law P is not necessarily known and should be estimated from the data. The most natural way is probably to estimate the law P by reading first all the data and to use this estimated law to encode the data. However this approach, called offline approach, suffers from two drawbacks: it requires to read first all the data, which leads to a delay, and to transmit the law, which leads to an overhead. A much better approach is to learn the law *online*, i.e. to estimate a law for each character using the ones previously seen. This strategy avoids the drawbacks of the offline one to the price of a law that can be less well estimated. Two major implementations of this stragegy exist: the first one is base on a straightforward modification of the arithmetic coding scheme and the second on a dictionary approach. For the arithmetic coding strategy, it suffices to notice that we use the law $P(w^k|w^1 \cdots w^{k-1})$ which is naturally learnt in an online way.

The dictionary approach differs from the explicit code strategy described so far. It is based on the online construction of a dictionary that is used to code new sequences of character using their position in the current dictionary. Numerous variations on the dictionary construction and its use exist (LZW,ZIP,...). Proofs of the asymptotic optimality of this strategy for stationary law exist.

6.5.2 GIF and PNG

GIF: This lossless image compression algorithm has been introduced by Compuserve in 1987. It is dedicated to 8 bits images seen as a list of pixels values (grayscale or colormap). A universal dictionary entropy encoder of the LZW family is used to encode those values.

PNG: This algorithm has been introduced in 1995 as replacement of GIF which was hindered by a patent complain. It is not meant to used for colormap images and can thus interpret pixel values as color intensities. Instead of coding the raw pixel values, it codes, using a (non patented) variant of universal dictionary coder, the difference between the pixel values and their predictions from previous values. It also adds some new functionality such as true color images and transparency.

Chapter 7

Lossy coding

7.1 Continuous signal and Distortion-Rate

Continuous dictionary and approximation: While the finite dictionary case can be easily extended to a countable setting, no direct extension can be made to a continuous dictionary $\mathcal{C}$: specifying a value requires an infinite amount of information. The only solution in such a setting is to allow error, i.e. code some approximation of the original values with a countable dictionary.

Quantifier and quantization error: a quantifier is an application $Q\mathcal{C} \to \mathcal{A}$ where $\mathcal{A}$ is a finite (or countable dictionary) of elements of $\mathcal{C}$. We measure the error between the value x and its quantized version $Q(x)$ by a loss ℓ so that the quantization error for x is given by $\ell(x, Q(x)$. To obtain a lossy code for $\mathcal{C}$, it suffices to provide a lossless code for $\mathcal{A}$.

Simplest case: the simplest quantifier for real values is the uniform scalar quantifier $Q : \mathbb{R} \to \Delta\mathbb{Z}$

$$x \mapsto \Delta \text{round}(x/\Delta)$$

and the simplest loss is $\ell(x, Q(x)) = (x - Q(x))^2$. Note that $\mathcal{A} = \Delta\mathbb{Z}$ can be identified with $\mathbb{Z}$.

Remark: the same strategy can be used if $\mathcal{C}$ is countable but very large...

Distortion-Rate: two different measures of the quality of such a coding scheme are used. The first one, the distortion, measure the average error

$$D = \int \ell(x, Q(x))p(x)\mathrm{d}x$$

while the second one, the (entropy) rate, measure the average of bits per symbol

$$\begin{aligned} R &= \sum_{a_k \in \mathcal{A}} l_k \int_{x, Q(x)=a_k} p(x)\mathrm{d}x \\ &\sim -\sum_{a_k \in \mathcal{A}} \int_{x, Q(x)=a_k} p(x)\mathrm{d}x \log_2 \int_{x, Q(x)=a_k} p(x)\mathrm{d}x. \end{aligned}$$

Those two quantities are expected to be *small* for a good quantifier but, obviously, a smaller distortion comes at a price of a larger rate and, conversely, a smaller rate implies a larger distortion. A good quantifier is one yielding a good tradeoff between those two quantities.

Scalar quantizer: We focus now on the case of a scalar quantifier $Q : \mathbb{R} \to \mathcal{A} = \{a_k\}_{k=1}^N$ with $a_k \in \mathbb{R}$ and $N \in \mathbb{N}^*$ (extension to $N = +\infty$ possible). The set $Q_k = \{x, Q(x) = a_k\}$ is called the **cell** associated to a_k. For sake of simplicity, we assume the error is measured with the **quadratic** loss $l(x, y) = (x - y)^2$. The distortion is then

$$D = \int (x - Q(x))^2 p(x)\mathrm{d}x = \sum_{k=1}^N \int_{x \in Q_k} (x - a_k)^2 p(x)\mathrm{d}x$$

while the rate is

$$R = -\sum_{k=1}^N \int_{x \in Q_k} p(x)\mathrm{d}x \log_2 \int_{x \in Q_k} p(x)\mathrm{d}x = -\sum_{k=1}^N p_{Q,k} \log_2 p_{Q,k}$$

if we let $p_{Q,k} = \int_{x \in Q_k} p(x)\mathrm{d}x$ (which depends only on Q_k).

Remarks:

- For a fixed set of cells (and thus a fixed R), the choice $a_k = \int_{x \in Q_k} xp(x)\mathrm{d}x$ leads to the minimal distortion D.
- For a fixed set of quantizer (but no fixed R), the choice $Q_k = \{x, \arg\min_{k'}(x - a_{k'})^2 = k\}$ minimizes the distortion.

Optimal quantization for a uniform source: we focus now on a simple case, the one of a uniform source belonging to $[0, 1]$, i.e. whose law has a density $p(x) = \mathbf{1}_{[0,1]}$ and study the choice of a quantifier Q with N cells. For any Q,

$$\begin{aligned} D &= \sum_{k=1}^N \int_{x \in Q_k} (x - \int_{u \in Q_k} u\mathrm{d}u)^2 \mathrm{d}x \\ R &= -\sum_{k=1} \int_{x \in Q_k} \mathrm{d}x \log_2 \int_{x \in Q_k} \mathrm{d}x = -\sum_{k=1}^N |Q_k| \log_2 |Q_k| \end{aligned}$$

where $|Q_k|$ is the measure of the cell Q_k.

Cell shape: Let N be fixed, assume $|Q_1|, \dots, |Q_k|$ are fixed, then

$$\begin{aligned} D &= \sum_{k=1}^{N} \int_{x \in Q_k} (x - \int_{u \in Q_k} u \mathrm{d}u)^2 \mathrm{d}x \\ &\geq \sum_{k=1}^{N} |Q_k| \frac{|Q_k|^2}{12} \mathrm{d}x \end{aligned}$$

with equality if and only if the cells are intervals. If we do not use entropy coding so that $R = -\log_2 N$ then the best choice is $|Q_k| = 1/N$ leading to $D = \frac{1}{12N^2}$.

Entropy coding: Let N be fixed

$$\begin{aligned} R &= -\sum_{k=1}^{N} |Q_k| \log_2 |Q_k| \\ &= -\frac{1}{2} \sum_{k=1}^{N} |Q_k| \log_2 |Q_k|^2 \\ &= -\frac{1}{2} \log_2 \sum_{k=1}^{N} |Q_k| |Q_k|^2 \\ &\geq -\frac{1}{2} \log_2 12 \sum_{k=1}^{N} |Q_k| \frac{|Q_k|^2}{12} \\ &\geq -\frac{1}{2} \log_2 12D \end{aligned}$$

with equality if $|Q_k| = 1/N$ and Q_k are intervals, so that the best choice is the same than the one without entropy coding.

Coding an uniform source: If $p(x) = \mathbf{1}_{[0,1]}$,

$$R \geq -\frac{1}{2} \log_2 12D \quad \Leftrightarrow \quad D \geq \frac{1}{12} 2^{-2R}$$

with equality if the quantizer is uniform of step $1/N$. In that case, i.e. uniform and equiprobable bins, we obtain the same result than with and without entropy coding:

$$R = \log N \quad \text{and} \quad D = \frac{1}{12N^2}.$$

Note that no explicit rules are given to attains intermediate rates or distortion...

High resolution assumption: to go beyond the uniform source case, we will rely on a classical approximation, the *high resolution assumption*. It is based on the observation that, locally, a pdf p can be approximated a constant $p(x_0)$ and thus, according to the previous analysis, locally the best choice is a uniform quantifier of size $\Delta(x_0)$... The high resolution assumption is a formalization of this idea as a limit property on a family of quantifier.

Limit quantifier width: More precisely, let $\Delta^N(x)$ be the width of the quantifier with N bins of this family, we assume that

$$N\Delta^N(x) \to \Delta(x).$$

Proposition 7.1: $\frac{1}{\Delta}$ is a density function.

Proof: By construction, $\Delta^N(x) \geq 0$ so that $\Delta(x) \geq 0$, now

$$\frac{1}{N}\sum_{k=2}^{N-1}\int_{Q_k}\frac{1}{\Delta^N(x)} = \frac{N-2}{N}$$

so that going to the limit yields the result.

$\frac{N}{\Delta(x)}$ can thus be interpreted as the local number of cells by units!

The high resolution assumption holds if

$$N^2D^N \to \int p(x)\frac{\Delta^2(x)}{12}dx = D^\infty$$
$$R^N - \log N \to -\int p(x)\log_2\left(p(x)\Delta(x)\right)dx = R^\infty.$$

Note that this is an assumption on both the quantifier family and the density law.

Practical use: replace D^N by D^∞/N^2 and R^n by $\log N + R^\infty$ and optimize in Δ. The resulting optimum should give an idea of the best possible quantifier...

Heuristic of high resolution assumption: We provide here without any justification a sequence of computation showing why the high resolution assumption may hold:

$$N^2D^N = N^2\int |x-Q(x)|^2p(x)$$
$$\sim \sum_{k=2}^{N-1}\int_{Q_k}|x-a_k|^2p(x)dx \sim \sum_{k=2}^{N-1}p(a_k)\frac{N^2\Delta^N(x)^3}{12}$$
$$\sim \sum_{k=2}^{N-1}p(a_k)\Delta^N(x)\frac{N^2\Delta^N(x)^2}{12} \sim \int p(x)\frac{\Delta(x)^2}{12}$$
$$R^N = -\sum_{k=1}^{N}\left(\int_{Q_k}p(x)dx\right)\log_2\left(\int_{Q_k}p(x)dx\right)$$
$$\sim -\sum_{k=2}^{N-1}\int_{Q_k}p(x)\log_2\left(\int_{Q_k}p(x')dx'\right) \sim -\sum_{k=2}^{N-1}\int_{Q_k}p(x)\log_2\left(p(x)\Delta^N(x)\right)$$
$$\sim -\sum_{k=2}^{N-1}\int_{Q_k}p(x)\left(\log_2\left(p(x)\Delta(x)\right) - \log_2 N\right)$$
$$\sim \log N - \int p(x)\log_2\left(p(x)\Delta(x)\right)$$

Best quantization without entropy coding: we use here a N cell quantifier without entropy coding so that $R^N = \log_2 N$ and try to optimize in Δ the approximate distortion $D^\infty N^2 = \int p(x)\frac{\Delta(x)^2}{12N^2}dx$ instead of the true one D^N. As $\int \frac{1}{\Delta(x)}dx = 1$, the Lagrangian formulation implies the existence of λ such that

$$\frac{\partial}{\mathrm{d}\Delta}\left(\int p(x)\frac{\Delta(x)^2}{12N^2}\mathrm{d}x + \lambda\left(\int \frac{1}{\Delta(x)}\mathrm{d}x - 1\right)\right) = 0.$$

This yields

$$\Delta(x) = \frac{\int p(x)^{1/3}\mathrm{d}x}{p(x)^{1/3}} \quad \Leftrightarrow \frac{1}{\Delta(x)} = \frac{p(x)^{1/3}}{\int p(x)^{1/3}\mathrm{d}x}.$$

This can be interpreted as a recommandation for smaller cell (or denser cell) where there is a high density of the source.

Summarizing the result in thid case yields

$$\begin{aligned} R &= \log_2 N \\ D &= \frac{1}{12N^2}\left(\int p(x)^{1/3}\mathrm{d}x\right)^3 = \frac{\left(\int p(x)^{1/3}\mathrm{d}x\right)^3}{12}2^{-2R} \end{aligned}$$

Best quantization with entropy coding: we assume now that we use an entropy coding scheme for the symbols so that we need to optimize $R^\infty + \log N$ and D^∞/N (instead of R^N and D^N). Some simple computations shows that

$$\begin{aligned} R^\infty &= -\int p(x)\log_2\left(p(x)\Delta(x)\right)\mathrm{d}x \\ &= H(p) - \int p(x)\log_2\Delta(x)\mathrm{d}x = H(p) - \frac{1}{2}\int p(x)\log_2\Delta(x)^2\mathrm{d}x \\ &\geq H(p) - \frac{1}{2}\log_2\int p(x)\Delta(x)^2\mathrm{d}x \sim H(p) - \frac{1}{2}\log_2 12D^\infty \end{aligned}$$

where $H(p) = -\int p(x)\log_2 p(x)\mathrm{d}x$. The inequality becomes an equality if and only if $\Delta(x)$ is constant (which is possible only if p is close to be compactly supported). This asymptotic distortion-rate analysis can be summarized by

$$R^\infty \geq H(p) - \frac{1}{2}\log_2 12D^\infty \quad \Leftrightarrow \quad D^\infty \geq \frac{e^{2H(p)}}{12}2^{-2R^\infty}$$

If we go back to $R^N \sim \log N + R^\infty$ and $D^N \sim \frac{D^\infty}{N^2}$, we obtain

$$R^N \gtrsim H(p) - \frac{1}{2}\log_2 12D^N \quad \Leftrightarrow \quad D^N \gtrsim \frac{e^{2H(p)}}{12}2^{-2R^N}$$

with *equality* when $\Delta(x)$ is constant. This leads to a gain with respect to the strategy without entropy coding because $e^{2H(p)} \leq \left(\int p(x)^{1/3}\mathrm{d}x\right)^3$.

Uniform quantization strategy: In practice, this suggests to use a uniform scalar quantifier with step Δ with an entropy coding scheme. If the high resolution assumption holds for this quanfifier then

$$D^N \sim D^{HR} = \frac{\Delta^2}{12}$$
$$R^N \sim R^{HR} = H(p) - \frac{1}{2}\log_2 12D^{HR} = H(p) - \log_2 \Delta.$$

Note that

$$R^{HR} = H(p) - \frac{1}{2}\log_2 12D^{HR} \quad \Leftrightarrow \quad D^{HR} = \frac{e^{2H(p)}}{12}2^{-2R^{HR}}$$

means

$$R^N \sim H(p) - \frac{1}{2}\log_2 12D^N \quad \Leftrightarrow \quad D^N \sim \frac{e^{2H(p)}}{12}2^{-2R^N}$$

One has to remember that the assumption may not hold!

7.1.1 Sub-band allocation

Sub-band allocation: we consider $x = (x^1, \ldots, x^B) \in \mathbb{R}^B$ and let p_i be the marginal law of x^i. We do not assume independence but will code those components independently. We want to study the optimal allocation in each of the different sub-bands. Namely, we assume we spend R_i bits to encode x^i (sub-band i) resulting in a distortion D_i so that the total distortion and the total rate are

$$D = \sum_{i=1}^{B} D_i \quad \text{and} \quad R = \sum_{i=1}^{B} R_i.$$

The question on how to allocates those bits to reach a given rate R or a given distortion D in an optimal way?

High-resolution analysis: We assume the high-resolution holds on all sub-bands so that the best choice is a **uniform quantifier** of bin length Δ_i

$$D_i^{HR} = \frac{\Delta_i^2}{12} \quad \text{and} \quad R_i^{HR} = H(p_i) - \log_2 \Delta_i.$$

The resulting total high resolution distortion and total high resolution rate are thus

$$D^{HR} = \sum_{i=1}^{B} \frac{\Delta_i^2}{12} \quad \text{and} \quad R^{HR} = \sum_{i=1}^{B} H(p_i) - \log_2 \Delta_i.$$

Assume we want to optimize Δ_i for a given (high resolution) rate R^{HR} then, using a Lagrangian formulation, it exists $\lambda \geq 0$ such that the optimal solutions satisfies

$$\frac{\partial}{\mathrm{d}\Delta_i}\left(\sum_{i=1}^{B} \frac{\Delta_i^2}{12} + \lambda \sum_{i=1}^{B} H(p_i) - \log_2 \Delta_i\right) = 0$$
$$\Leftrightarrow \frac{\Delta_i}{6} - \lambda \frac{1}{\Delta_i \log 2} = 0.$$

This yields $\Delta_i = \Delta$ and thus

$$R^{HR} = \sum_{i=1}^{B} H(p_i) - B \log_2 \Delta.$$

We deduce

$$\Delta = \sqrt{\frac{12D}{B}} = 2^{\frac{1}{B}\left(\sum_{i=1}^{B} H(p_i) - R^{HR}\right)}.$$

Plugging this into the expression of D^{HR} and R^{HR} yields

$$D^{HR} = \frac{Be^{\frac{2}{B}\sum_{i=1}^{B} H(p_i)}}{12} 2^{-2\frac{R^{HR}}{B}} \quad \text{and} \quad R^{HR} = \sum_{i=1}^{B} H(p_i) - \frac{B}{2} \log_2 \frac{12D^{HR}}{B}.$$

Provided the high resolution assumption holds for the uniform quantifier, we deduce that the optimal choice is use the same bin width Δ for all sub-band and that the resulting quantifier satisfies

$$D \sim D^{HR} = B\frac{\Delta^2}{12} \quad \text{and} \quad R \sim R^{HR} = \sum_{i=1}^{B} H(p_i) - B \log_2 \Delta$$

so that

$$D \sim \frac{Be^{\frac{2}{B}\sum_{i=1}^{B} H(p_i)}}{12} 2^{-2\frac{R}{B}} \quad \text{and} \quad R \sim \sum_{i=1}^{B} H(p_i) - \frac{B}{2} \log_2 \frac{12D}{B}.$$

Weighted Distortion optimization: when the distortion is measured with different weights for each sub-band, the previous analysis is only slightly modified. If we define the weighted high resolution distorsion and the high distortion rate by

$$D_W^{HR} = \sum_{i=1}^{B} W_i \frac{\Delta_i^2}{12} \quad \text{and} \quad R^{HR} = \sum_{i=1}^{B} H(p_i) - \log_2 \Delta_i,$$

then, again using the Lagragangian formulation, we obtain that the optimal choice satifies $\Delta_i = \frac{\Delta}{\sqrt{W_i}}$ and thus

$$D_W^{HR} = B\frac{\Delta^2}{12} \quad \text{and} \quad R^{HR} = \sum_{i=1}^{B} (H(p_i) - \frac{1}{2}\log_2 W_i) - B \log_2 \Delta$$

for a suitable Δ. One obtains thus

$$D_W^{HR} = \frac{Be^{\frac{2}{B}\sum_{i=1}^{B} (H(p_i) - \frac{1}{2}\log_2 W_i)}}{12} 2^{-2\frac{R^{HR}}{B}}$$
$$\text{and} \quad R = \sum_{i=1}^{B} (H(p_i) - \frac{1}{2}\log_2 W_i) - \frac{B}{2} \log_2 \frac{12D_W^{HR}}{B}$$

Provided the high resolution assumption holds for the uniform quantifier in all sub-bands, this implies that the best choice is to use a uniform quantifier of size $\frac{\Delta}{\sqrt{W_i}}$ in each sub-band and that

$$D_W \sim D_W^{HR} = B\frac{\Delta^2}{12} \quad \text{and} \quad R \sim R^{HR} = \sum_{i=1}^{B}(H(p_i) - \frac{1}{2}\log_2 W_i) - B\log_2 \Delta$$

so that

$$D_W \sim \frac{Be^{\frac{2}{B}\sum_{i=1}^{B}(H(p_i)-\frac{1}{2}\log_2 W_i)}}{12}2^{-2\frac{R}{B}} \quad \text{and} \quad R \sim \sum_{i=1}^{B} H(p_i) - \frac{B}{2}\log_2 \frac{12D_W}{B}.$$

7.1.2 Transform coding

Basis and transform coding: assume now, we observe $f \in V$, a (hilbertian) space of dimension B. In any basis $(b_i)_{i=1}^B$,

$$f = \sum_{i=1}^{B}\langle f, \widetilde{b}_i\rangle b_i$$

where $(\widetilde{b}_i)_{i=1}^B$ is the dual basis. f can thus be described by the finite sequence of coefficients $(\langle f, b_i\rangle)_{1\leq i\leq B}$. In a transform coding scheme, those coefficients are then coded in a lossy scheme consisting of a quantization step followed by an entropy coding step.

Example: f is an image described in the canonical basis and the quantization step amount to a *reduction* of the number of color used.

Independent quantification: For sake of simplicity, we assume we use B independent quantifiers $(Q_i)_{i=1}^B$ for the B coefficients and let

$$f_Q = \sum_{i=1}^{B} Q_i(\langle f, \widetilde{b}_i\rangle)b_i.$$

Distortion/Quantization error: For sake of simplicity, we assume that the quantization error for f is measured in term of the l^2 error of the coefficients so that

$$\ell(f, f_Q) = \sum_{i=1}^{B} W_i \left|\langle f, \widetilde{b}_i\rangle - Q_i(\langle f, \widetilde{b}_i\rangle)\right|^2.$$

Note that if $(b_i)_{i=1}^B$ is an orthonormal basis then $\widetilde{b}_i = b_i$ and for $W_i = 1$ $\ell(f, f_Q) = \|f - f_Q\|^2$.

Basis and high resolution assumption: If we assume that the high resolution assumption holds for the uniform quantifier for all coefficients then the best strategy is to use a uniform quantifier of step $\frac{\Delta}{\sqrt{W_i}}$. The previous rate/distortion analysis yields then

$$D_W \sim B\frac{\Delta^2}{12} \quad \text{and} \quad R \sim \sum_{i=1}^{B}\left(H(\langle f, \widetilde{b}_i\rangle) - \frac{1}{2}\log_2 W_i\right) - B\log_2 \Delta$$

which implies

$$D_W \sim \frac{Be^{\frac{2}{B}\sum_{i=1}^{B}\left(H(\langle f,\widetilde{b_i}\rangle)-\frac{1}{2}\log_2 W_i\right)}}{12}e^{-2\frac{R}{B}}$$
$$\text{and} \quad R \sim \sum_{i=1}^{B}\left(H(\langle f,\widetilde{b}_i\rangle)-\frac{1}{2}\log_2 W_i\right)-\frac{B}{2}\log_2\frac{12D_W}{B}.$$

Orthonormal bases: for any orthonormal basis $(b_i)_{i=1}^B$ and the choice $W_i = 1$, the distortion is measured with respect to the same L^2 norm. The best strategy is to use the same bin width Δ for all coefficients and leads to the relationships

$$D \sim B\frac{\Delta^2}{12} \quad \text{and} \quad R \sim \sum_{i=1}^{B} H(\langle f, b_i\rangle) - B\log_2\Delta$$

which implies

$$D \sim \frac{Be^{\frac{2}{B}\sum_{i=1}^{B}H(\langle f,b_i\rangle)}}{12}e^{-2\frac{R}{B}} \quad \text{and} \quad R = \sum_{i=1}^{B} H(\langle f, b_i\rangle) - \frac{B}{2}\log_2\frac{12D}{B}.$$

As the loss is the same whatever the orthonormal basis, the question of the basis choice in such a coding scheme has a sense. A straightforward observation shows that the best basis is the one minimizing

$$\sum_{i=1}^{B} H(\langle f, b_i\rangle).$$

Best basis for Gaussian process: in this paragraph, we assume that f is a centered Gaussian process: $f \in \mathbb{R}^B \sim \mathcal{N}(0,\Gamma)$ and study the best orthornormal basis.

Theorem 7.1: As long as the high-resolution assumption holds, the best orthonormal basis is the one that diagonalizes Γ.

Proof: The proof is based on an explicit formula for $\sum_{i=1}^{B} H(\langle f, b_i\rangle)$

Note that this result can be understand intuitively. The diagonalization basis is the one that decorrelates the coefficients and thus they are independent. There is thus no loss by coding them independently as we are doing here. Finally, this theorem justify the use of Fourier type basis for locally stationary process as we know that such a basis diagonalizes the *covariance matrix*...

Sparse coding: For sake of simplicity, we assume now that the b_i are orthonormal basis and that we use an independent lossy coding scheme for each coefficient. We have thus

$$f = \sum_{i=1}^{B}\langle f, b_i\rangle b_i$$
$$f_Q = \sum_{i=1}^{B} Q(\langle f, b_i\rangle) b_i$$

and the quantization error is given by

$$\|f - f_Q\|^2 = \sum_{i=1}^{B} |\langle f, b_i\rangle - Q(\langle f, b_i\rangle)|^2.$$

In practice, one observe that, for *suitable* bases and for *interesting* signal, a lot of coefficients are very small, much smaller than $\Delta/2$ for reasonable Δ and thus quantized to 0. Such signals are called sparse and require a different coding analysis as the high resolution assumption does not hold for the 0 bin.

Sparse quantification error: we focus first on a deterministic upper bound of the quantification error

$$\begin{aligned} D(\Delta) = \|f - f_Q\|^2 &= \sum_{i=1}^{B} |\langle f, b_i\rangle - Q(\langle f, b_i\rangle)|^2 \\ &= \sum_{1\leq i\leq B, |\langle f,b_i\rangle|\leq \Delta/2} |\langle f, b_i\rangle|^2 + \sum_{1\leq i\leq B, |\langle f,b_i\rangle|>\Delta/2} |\langle f, b_i\rangle|^2 \\ &\leq \underbrace{\sum_{1\leq i\leq B, |\langle f,b_i\rangle|\leq \Delta/2} |\langle f, b_i\rangle|^2}_{\text{approximation error}} + \underbrace{\frac{\Delta^2}{4} |\{1 \leq i \leq B, |\langle f, b_i\rangle| > \Delta/2\}|}_{\text{quantization error}} \\ &\leq A(\Delta) + \frac{\Delta^2}{4} M(\Delta) \end{aligned}$$

The approximation error $A(\Delta)$ can be much smaller than $\frac{\Delta^2}{4} |\{1 \leq i \leq B, |\langle f, b_i\rangle| \leq \Delta/2\}| = \frac{\Delta^2}{4}(B - M(\Delta))$ and thus $\|f - f_Q\|^2$ can be much smaller than $B\frac{\Delta^2}{4}$.

Sparse rate: we describe now a two step sparse coding strategy: first code for each coefficient if it is quantified to 0 or not, second code the $M(\Delta)$ non zero coefficients. For sake of simplicity, we assume first that we don't use a entropy coder. To code the position, we can code $M(\Delta)$ using at most $\log_2 B$ bits and then the position of the $M(\Delta)$ coefficients by $\log_2 \binom{B}{M(\Delta)}$ bits. It remains then to code the $M(\Delta)$ values with $\log_2(2\max) - \log_2 \Delta$. The total length of this code is thus

$$R(\Delta) = \log B + \log_2 \binom{B}{M(\Delta)} + M(\Delta)\left(\log_2(2\max) - \log_2 \Delta\right).$$

Using now $\binom{B}{M(\Delta)} \leq 2^{BH(M(\Delta)/B)}$, we deduce

$$R(\Delta) \leq \log B + BH(M(\Delta)/B) + M(\Delta)\left(\log_2(2\max) - \log_2 \Delta\right).$$

We have thus

$$\begin{aligned} D(\Delta) = \|f - f_Q\|^2 &\leq A(\Delta) + \frac{\Delta^2}{4} M(\Delta) \\ R &\leq \log B + BH(M(\Delta)/B) + M(\Delta)\left(\log_2(2\max) - \log_2 \Delta\right). \end{aligned}$$

Assume now that Δ is large (at least larger than $\Delta_{\min}$), so that $M(\Delta)$ is small with respect to B, we can provide a simpler estimate of R:

$$\begin{aligned} R(\Delta) &\leq \log B - B\left(M(\Delta)/B \log_2(M(\Delta)/B) + (1 - M(\Delta)/B)\log_2(1 - M(\Delta/B))\right) \\ &\quad + M(\Delta)\left(\log_2(2\max) - \log_2 \Delta\right) \\ &\lesssim M(\Delta)\left(1 - \log_2(M(\Delta)/B) + \log_2(2\max/\Delta_{\min})\right) \\ &\lesssim M(\Delta)\left(1 + \log_2 B + \log_2(2\max/\Delta_{\min})\right). \end{aligned}$$

Proof: As

$$n^n = (k + (n-k))^n \geq \binom{n}{k} k^k (n-k)^{n-k},$$

we deduce

$$\begin{aligned} \binom{n}{k} &\leq \frac{n^n}{k^k(n-k)^{n-k}} = \left(\frac{1}{(k/n)^{k/n}(1-k/n)^{1-k/n}}\right)^n = \left(\frac{1}{2^{-H(k/n)}}\right)^n \\ \binom{n}{k} &\leq 2^{nH(k/n)} \end{aligned}$$

Compressibility: If there exist $\gamma \in [0,1]$ such that

$$A(\Delta) + M(\Delta)\frac{\Delta^2}{4} \leq CB\Delta^{2\gamma} \ll B\frac{\Delta^2}{4}$$

for $\Delta \geq \Delta_{\min}$, we say that the signal is compressible.

In that case,

$$M(\Delta) \leq 4CB\Delta^{2(\gamma-1)}$$

which implies

$$\begin{aligned} R(\Delta) &\lesssim 4CB\left(1 + \log_2 B + \log_2(2\max/\Delta_{\min})\right)\frac{1}{\Delta^{2(1-\gamma)}} \\ &\lesssim 4CB\left(1 + \log_2 B + \log_2(2\max/\Delta_{\min})\right)\left(\frac{CB}{D(\Delta)}\right)^{\frac{1-\gamma}{\gamma}} \\ R(\Delta) &\lesssim 4(CB)^{\frac{1}{\gamma}}\left(1 + \log_2 B + \log_2(2\max/\Delta_{\min})\right) D(\Delta)^{-\frac{1-\gamma}{\gamma}} \end{aligned}$$

and thus

$$D(\Delta) \lesssim \left(4\left(1 + \log_2 B + \log_2(2\max/\Delta_{\min})\right)\right)^{\frac{\gamma}{1-\gamma}} (CB)^{\frac{1}{1-\gamma}} R(\Delta)^{-\frac{\gamma}{1-\gamma}}.$$

We do not obtain the exponential decay of the high resolution analysis $D(\Delta) \sim 2^{-R(\Delta)}$ which seems less attractive but, we are interested in low bit rate, that is the behavior for small R (but not too small as R can barely go to 0 here). In that case, one observes that the decay of the error is much faster with the sparse coding strategy than with the high resolution one.

7.2 JPEG, block transform and MP3

JPEG: this image compression *standard* has been proposed in 1990 by an expert committee whose members were coming from both industry and academy. It relies on a transform coding strategy. The image is split into 8×8 blocks that are further transformed by a 2D Fourier type basis (Discrete Cosine Transform). The resulting coefficients are then quantized and entropy coded using a different strategy for the means of the blocks than for the other coefficients. The first ones are first predicted by the value of the mean of the previous block and then encoded by an Huffman code. The second ones are encoded block by block through a combination of Run Length Encoding and Laplacian model. The quantification strategy is adapted to the eye perceptual properties: the high frequencies are more roughly quantized than the low ones.

Block transforms: more generaly, the block strategy can be used to generate basis of $l^2(\mathbb{Z})$ from finite basis of $\ell^2(\{0,\dots,N\})$. It suffices to notice that for any increasing sequence a_k such that $\lim_{k\to-\infty} a_k = -\infty$ and $\lim_{k\to\infty} a_k = \infty$ we have $\mathbb{Z} = \cup_{k\in\mathbb{Z}}\{a_k,\dots,a_{k+1}-1\}$. We deduced immediately that if $(b_{k,l})_{0\leq l<a_{k+1}-a_k}$ is an orthonormal basis of $\ell^2(\{a_k,\dots,a_{k+1}-1\})$ then $(b_{k,l})_{k\in\mathbb{Z},0\leq l<a_{k+1}-a_k}$ is an orthonormal basis of $\ell^2(\mathbb{Z})$. Given a family $\{\{e_l^N\}_{0\leq l<N}\}_{N\in\mathbb{N}}$ of local orthonormal basis of $\ell^2(\{0,\dots,N\})$, we obtain that

$$\bigcup_{k\in\mathbb{Z}} \left\{ e_l^{a_{k+1}-a_k}[\cdot - a_k] \right\}_{0\leq l<a_{k+1}-a_k}$$

is a valid block by block basis of $\ell^2(\mathbb{Z})$.

Note that a similar construction can be performed in the continuous domain.

Local basis family: a classical choice of local orthonormal basis is to used a local Fourier type basis. The easiest choice is to use the Discrete Fourier Transform

$$e_l^N[n] = \frac{1}{\sqrt{N}} e^{i\frac{2\pi}{N}ln}.$$

This solution implies an implicit periodization and thus suffer from strong discontinuity artifact at the boundaries of the block. A better solution is to used a local Cosine basis in which the periodization is preceded by a symmetrization, so that only derivative are discontinuous at the boundaries. The most classical choice is the Discrete Cosine Transform I:

$$e_l^N[n] = \lambda_l \sqrt{\frac{2}{N}} \cos\left(\frac{\pi}{N} l(n+\frac{1}{2})\right)$$

for which a Fast transform is available. This is the one used in the JPEG standard.

Overlapped block transform: the previous local basis strategy still suffers from discontinuity artefacts, as the reconstruction may be discontinuous at boundaries. The overlapped block strategy is an overlapping windowing strategy proposed to reduce this issue. It relies on a clever orthonormal projection operator, defined from a suitable window family, the signal in a direct sum of spaces V_k space of size $a_{k+1}-a_k$ located in a neighborhood of $\{a_k\dots,a_{k+1}-1\}$. Those projections are specified by their coefficients in a suitable basis of the V_k, with a systematic

construction from any basis of $\{a_k \dots, a_{k+1} - 1\}$. The most used choice is the Discrete Cosine Transform IV

$$b_{l,k}[n] = g_k[n]\sqrt{\frac{2}{a_{k+1} - a_k}} \cos\left(\pi(l + 1/2)\frac{n - a_k + 1/2}{a_{k+1} - a_k}\right)$$

with g_k the window of space V_k. A fast transform is available and is used in the MP3 standard.

MP3: the compression method relies on all the tools seen so far. It decomposes the sound in projection in spaces of adapted sizes, compute the cosine type transform coefficients for each window, use a psychoaccoustic model to quantify the coefficients, capitalizing on the different sensitivities for different frequencies and on the masking effect of a strong frequency, and then encode the coefficients with an entropy coder.

Bibliography

[1] Peter Brockwell and Richard Davis. *Time series: Theory and Methods*. Springer-Verlag, 1991. ISBN: 0387974296.

A comprehensive book on Time Series including WSSP.

[2] Paul L. Butzer and Rudolf L. Stens. "Sampling theory for not necessarily band-limited functions: A historical review". In: *SIAM review* 34.1 (1992), pp. 40–53.

A historical review on sampling theory detailling its history from Lagrange polynomial interpolation to subtle extension of Shannon theorem

[3] Maurizio Ciampa, Marco Franciosi, and Mario Poletti. "A note on impulse response for continuous, linear, time-invariant, continuous-time systems". In: *IEEE Transactions on Circuits and Systems* 53.1 (2006), pp. 106–113.

A very well written article debunking the myth that a LTI system is always a convolution system.

[4] Thomas Cover and Joy Thomas. *Elements of Information Theory*. 2nd ed. Wileys, 2006. ISBN: 0471241954.

The reference on Information theory which provides a comprehensive treatement of the theoretical aspects.

[5] Ingrid Daubechies. *Ten lectures on Wavelets*. 1st ed. Society for Industrial and Applied Mathematics, 1992. ISBN: 0898712742.

A book dedicated to Time Frequency analysis with an emphasis on wavelets which remains a reference on the subject.

[6] Claude Gasquet and Patrick Witomski. *Fourier Analysis and Applications. Filtering, Numerical Computation, Wavelets*. 1st ed. Springer, 1999. ISBN: 978-1-4612-1598-1.

A wonderful book on Fourier Analysis also available in French (Analyse de Fourier et Applications. Filtrage, calcul numérique et ondelettes chez Dunod). It covers basic Fourier Analysis as well as the distribution setting and gives comprehensive proofs.

[7] John Makhoul. "Linear Prediction: A Tutorial Review". In: *Proceedings of the IEEE* 63.4 (Apr. 1975), pp. 561–580.

A review of Linear Prediction explaining the Least Square approach in both the determinstic and the random modeling case as well as its spectral interpretation.

[8] Stéphane Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. 3rd ed. Academic Press, 2008. ISBN: 0123743702.

The reference on wavelet signal processing and sparse signal analysis.

[9] Stéphane Mallat. *Traitement du Signal*. École Polytechnique, 1999.

The previous poly from École Polytechnique from which this one is largely inspired.

[10] Lawrence R. Rabiner and Ronald W. Schafer. “Introduction to Digital Speech Processing”. In: *Foundations and Trends in Signal Processing* 1.1-2 (2007), pp. 1–194.
A well written survey on Digital Speech Processing encompassing voice modeling, LPC, vocoder and more.

[11] Laurent Schwartz. *Théorie des distributions*. Reprint of the 1966 edition. Hermann, 1997.
A reprint of the original book from L. Schwartz written in 1966, which remains one of the best references on distributions.

[12] Eric Stade. *Fourier Analysis*. Wiley, 2005. ISBN: 0471669849.
A Fourier Analysis book pleasantly written by a number theoretician which puts an emphasis on applications and on numerical analysis.

[13] Claudio Weidmann and Martin Vetterli. “Rate Distortion Behavior of Sparse Sources”. In: *IEEE Transactions on Information Theory* 58.8 (2012), pp. 4969–4992.
A recent article studying the sparse coding case from the mathematical point of view