Lecture 2

Christa Cuchiero
based on joint lectures with M. Gambera, J. Teichmann, H. Wutte

Institute of Statistics and Mathematics
Vienna University of Economics and Business

Part I

Wavelet frames (for deep neural networks)

A brief history of wavelet analysis

- Wavelets have their origins in signal analysis and engineering.
- The term "wavelet" was coined in the 1980s in geophysics by Jean Morlet, Alex Grossman for functions that generalize the short-term Fourier transform.
- In the 1990s, a veritable wavelet boom arose, triggered by
  - the discovery of compact, continuous (to any order of differentiability) and orthogonal wavelets by Ingrid Daubechies (1988) and
  - the development of the algorithm of fast wavelet transformation (FWT) using multi-scale analysis (MRA) by Stéphane Mallat and Yves Meyer (1989).

The what, the why and the how of wavelets?

The following introduction to wavelets is based on Ingrid Daubechies' "Ten lectures on wavelets", Chapter 1:

- Wavelets provide a tool for time-frequency localization.
- Given a signal $f(t)$ (we here assume for simplicity that t is a continuous variable and f a function in one variable), one is interested in its frequency content locally in time.
- The wavelet transform of a signal evolving in time depends therefore on two variables: frequency and time.

Time frequency localization

- The standard Fourier transform,

$$\mathcal{F}f(\omega) = \frac{1}{\sqrt{2\pi}} \int e^{-\mathrm{i}\omega t} f(t)dt$$

gives a representation of the frequency content of f, but information concerning time-localization cannot be read off easily from $\mathcal{F}f$.

- Time-localization can be achieved by a windowed Fourier transform, i.e.

$$\mathcal{F}^{\mathrm{win}} f(\omega, t) = \frac{1}{\sqrt{2\pi}} \int e^{-\mathrm{i}\omega s} f(s) g(s-t) ds.$$

- In its discrete version t and ω are assigned regularly spaced values: $t = nt_0, \omega = m\omega_0$, for $n, m \in \mathbb{Z}$ and $\omega_0, t_0 > 0$ fixed:

$$\mathcal{F}^{\mathrm{win}}_{m,n} f = \frac{1}{\sqrt{2\pi}} \int e^{-\mathrm{i}m\omega_0 s} f(s) g(s - nt_0) ds.$$

The windowed Fourier transform

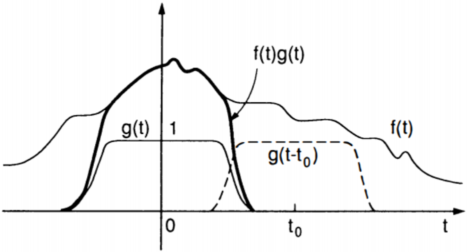

FIG. 1.1. *The windowed Fourier transform: the function $f(t)$ is multiplied with the window function $g(t)$, and the Fourier coefficients of the product $f(t)g(t)$ are computed; the procedure is then repeated for translated versions of the window, $g(t - t_0)$, $g(t - 2t_0)$, $\cdots$.*

The wavelet transform

- The continuous and discrete wavelet transform formulas are analogous to the continuous and discrete windowed Fourier, i.e.

$$T^{\text{wav}} f(a,b) = \frac{1}{\sqrt{|a|}} \int f(t)\psi(\frac{t-b}{a})dt$$

and

$$T^{\text{wav}}_{m,n} f = \frac{1}{\sqrt{|a_0|^m}} \int f(t)\psi(a_0^{-m}t - nb_0)dt$$

- The function ψ is sometimes called mother wavelet and satisfies $\int \psi(t)dt = 0$.
- $T^{\text{wav}}_{m,n} f$ is again obtained from $T^{\text{wav}} f(a,b)$ by restricting a, b to only discrete values: $a = a_0^m$, $b = nb_0 a_0^m$ in this case, with m, n ranging over $\mathbb{Z}$, and $a_0 > 1, b_0 > 0$ fixed.

Analogies/differences to the windowed Fourier transform

- **Similarity:** both transforms take the inner products of f with a family of functions indexed by two labels: $g^{\omega,t}(s) = e^{i\omega s} g(s-t)$ in the Fourier case and

$$\psi^{a,b}(s) = \frac{1}{\sqrt{|a|}} \psi(\frac{s-b}{a}).$$

  in the wavelet case. The functions $\psi^{a,b}$ are called wavelets

- **Difference:** shapes of $g^{\omega,t}$ and $\psi^{a,b}$.
  - The functions $g^{\omega,t}$ all consist of the same g, translated to the proper time location, and "filled in" with higher frequency oscillations. All the $g^{\omega,t}$ have the same width.
  - In contrast, the $\psi^{a,b}$ have time-widths adapted to their frequency: high frequency $\psi^{a,b}$ are very narrow, while low frequency $\psi^{a,b}$ are much broader. As a result, the wavelet transform is better able than the windowed Fourier transform to "zoom in" on very short lived high frequency phenomena.

Comparison between windowed Fourier transform

- A typical choice for ψ is $\psi(t) = (1 - t^2)\exp(-t^2/2)$, sometimes called the mexican hat function, illustrated below.

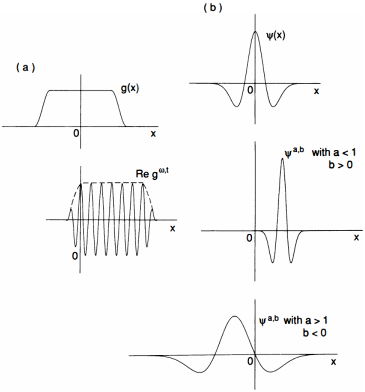

FIG. 1.2. *Typical shapes of* (a) *windowed Fourier transform functions* $g^{\omega,t}$, *and* (b) *wavelets* $\psi^{a,b}$. *The* $g^{\omega,t}(x) = e^{i\omega x} g(x-t)$ *can be viewed as translated envelopes* g, *"filled in" with higher frequencies; the* $\psi^{a,b}$ *are all copies of the same functions, translated and compressed or stretched.*

Different types of wavelet transforms

A The continuous wavelet transform $T^{\text{wav}} f(a, b)$ and

B The discrete wavelet transform $T^{\text{wav}}_{m,n} f$

Different types of wavelet transforms

A The continuous wavelet transform $T^{\text{wav}} f(a, b)$ and

B The discrete wavelet transform $T^{\text{wav}}_{m,n} f$

Within the discrete wavelet transform we distinguish further between

B1 Redundant discrete systems, so-called wavelet frames and

B2 Orthonormal (and other) bases of wavelets (in $L^2(\mathbb{R})$)

A: the continuous wavelet transform

- A function can be reconstructed from its wavelet transform by means of the "resolution of identity" formula,

$$f = C_{\psi}^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{a^2} \langle \psi^{a,b}, f \rangle \, \psi^{a,b} da \, db$$

  where $\langle \cdot, \cdot \rangle$ denotes the L^2 scalar product.

- The constant C_{ψ}, depends only on ψ and is given by

$$C_{\psi} = 2\pi \int_{-\infty}^{\infty} \widehat{\psi}(\xi)\xi^{-1} d\xi$$

  and we assume $C_{\psi} < \infty$. If ψ is in $L^1(\mathbb{R})$ (this is the case in all examples of practical interest) , then $\widehat{\psi}$ is continuous, so that C_{ψ}, can be finite only if $\widehat{\psi}(0) = 0$, i.e. $\int \psi(t)dt = 0$.

- The above formula can viewed as a way to write f as a superposition of wavelets where the coefficients in this superposition are given by the wavelet transform of f.

B1: the discrete but redundant wavelet transform-frames.

- The dilation parameter a and the translation parameter b both take only discrete values.
- For a we choose the integer (positive and negative) powers of one fixed dilation parameter $a_0 > 1$, i.e. , $a = a_0^m$.
- Different values of m correspond to wavelets of different widths.
- The discretization of b should thus depend on m: narrow (high frequency) wavelets are translated by small steps, while wider (lower frequency) wavelets are translated by larger steps.
- Since the width of $\psi(a_0^{-m}x)$ is proportional to a_0^m, we choose therefore to discretize b by $b = nb_0 a_0^m$
- The corresponding discretely labelled wavelets are therefore

$$\psi_{m,n} = a_0^{-m/2}\psi(a_0^{-m}(x - nb_0 a_0^m))) = a_0^{-m/2}\psi(a_0^{-m}x - nb_0)$$

B1: Recovering f?

- In the discrete case there does not exists a "resolution of the identity" formula as in the continuous case.
- Reconstruction of f from $(T^{\text{wav}}_{m,n}(f))_{m,n}$, if at all possible, must therefore be done by some other means.
- The following questions naturally arise:
  - Is it possible to characterize f completely by knowing $(T^{\text{wav}}_{m,n}(f))_{m,n}$, or in other words
  - Can any function be written as a superposition of $\psi_{m,n}$?
- As in the continuous case, these discrete wavelet transforms often provide a very redundant description of the original function.

B2: Orthonormal wavelet bases

- For some very special choices of ψ and a_0, b_0, the $\psi_{m,n}$ constitute an orthonormal basis for $L^2(\mathbb{R})$.
- In particular, if we choose $a_0 = 2$, $b_0 = 1$, then there exist ψ with good time-frequency localization properties, such that the

$$\psi_{m,n}(x) = 2^{-m/2}\psi(2^{-m}x - n)$$

  constitute an orthonormal basis for $L^2(\mathbb{R})$.
- The oldest example of a function ψ for which the $\psi_{m,n}$ as defined above constitute an orthonormal basis for $L^2(\mathbb{R})$ is the Haar function.

$$\psi(x) = \begin{cases} 1, & 0 \leq x < \frac{1}{2}, \\ -1, & \frac{1}{2} \leq x < 1, \\ 0, & \text{else} \end{cases}$$

Frames

We return to the setting of $B1$ where we do not necessarily have a basis but rather redundant wavelet frames.

Definition

A family of functions $(g_j)_{j\in J}$ in a Hilbert space H and J a countable index set is called a frame if there exist $A > 0, B < \infty$ so that, for all $f \in H$,

$$A\|f\|^2 \leq \sum_{j\in J} |\langle f, g_j\rangle|^2 \leq B\|f\|^2$$

We call A and B the frame bounds. If the two frame bounds are equal, $A = B$, then the frame is called a tight frame, i.e. we have $\sum_{j\in J} |\langle f, g_j\rangle|^2 = A\|f\|^2$.

Frames versus (orthonormal) bases

Proposition

If $(g_j)_{j\in J}$ is a tight frame, with frame bound $A = 1$, and if $\|g_j\|^2 = 1$ for all $j \in J$, then the g_j constitute an orthonormal basis.

- In general, frames and even tight frames are however not orthonormal bases.
- Nevertheless at least for tight frames the function can expressed similarly. Indeed, by the polarization identity we have

$$A\langle f, h\rangle = \sum_j \langle f, g_j\rangle\langle h, g_j\rangle$$

so that $f = A^{-1}\sum_j \langle f, g_j\rangle g_j$ at least weakly.

Frame operator

How does it work for general frames?

Definition

If $(g_j)_{j\in J}$ is a frame in H, then the frame operator F is the linear operator from H to $\ell^2(J) = \{(c_j)_{j\in J} \mid \sum |c_j|^2 < \infty\}$ defined by $(Ff)_j = \langle f, g_j \rangle$.

- It follows from the frame bounds that $\|Ff\|^2 \leq B\|f\|^2$. It is thus a bounded operator.
- The adjoint F^* of F is easy to compute:

$$\langle F^* c, f\rangle = \langle c, Ff\rangle = \sum_j c_j \langle g_j, f\rangle$$

Hence $F^* c = \sum_j c_j g_j$.

Frame operator

The definition of the frame operator implies further that

$$\sum_j |\langle f, g_j \rangle|^2 = \|Ff\|^2 = \langle F^* Ff, f \rangle.$$

Hence, by the frame bounds we have

$$A\,\mathrm{Id} \leq F^* F \leq B\,\mathrm{Id}$$

This implies, in particular, that $F^* F$ is invertible, by the following lemma.

Lemma

If a nonnegative bounded linear operator S on H is bounded from below by a strictly positive constant α, then operator is invertible and its inverse S^{-1} is bounded from above by α^{-1}.

Towards dual frames

- From the above we have that $(F^*F)^{-1}$ is well defined and $\|(F^*F)^{-1}\| \leq A^{-1}$.
- Applying the operator $(F^*F)^{-1}$ to the vectors g_j leads to an interesting new family of vectors, which we denote by $\widetilde{g}_j = (F^*F)^{-1}g_j$.

Proposition

The $(\widetilde{g}_j)_{j\in J}$ constitute a frame with frame bounds B^{-1} and A^{-1}, i.e.,

$$B^{-1}\|f\|^2 \leq \sum_{j\in J} |\langle f, \widetilde{g}_j\rangle|^2 \leq A^{-1}\|f\|^2$$

The associated frame operator $\widetilde{F} : H \to \ell^2(J)$, $(\widetilde{F}f)_j = \langle f, \widetilde{g}_j\rangle$ satisfies $\widetilde{F} = F(F^*F)^{-1}$, $\widetilde{F}^*\widetilde{F} = (F * F)^{-1}$, $\widetilde{F}^*F = Id = F^*\widetilde{F}$ and $F\widetilde{F}^* = \widetilde{F}F^*$ is the orthogonal projection operator, in $\ell^2(J)$ onto $\mathrm{Ran}(F) = \mathrm{Ran}(\widetilde{F})$.

Dual frames

We call $(\widetilde{g}_j)_{j\in J}$ the dual frame of $(g_j)_{j\in J}$. The conclusions of the above Proposition namely $\widetilde{F}^*F = \mathrm{Id} = F^*\widetilde{F}$ means

$$\widetilde{F}^*Ff = \sum_j \langle f, g_j\rangle \widetilde{g}_j = f = F^*\widetilde{F}f = \sum_j \langle f, \widetilde{g}_j\rangle g_j.$$

- This means that we have a reconstruction formula for f from the $\langle f, g_j\rangle$.
- At the same time we have also obtained a recipe for writing f as a superposition of g_j.

"Optimality" of Dual frames

Frames, even tight frames, are generally not (orthonormal) bases because the g_j are typically not linearly independent. This means that for a given f, there exist many different superpositions of the g_j which all add up to f. The above superposition is nevertheless special in the following sense.

Proposition

If $f = \sum_j c_j g_j$ *for some* $(c_j) \in \ell^2(J)$ *and if not all* $c_j = \langle f, \widetilde{g}_j \rangle$, *then* $\sum_j |c_j|^2 > \sum_j |\langle f, \widetilde{g}_j \rangle|^2$.

Wavelet frames

- Wavelet frames can be construced from a family of averaging kernels ("father functions" or scaling functions) satisfying certain conditions.
- This is connected to so-called multiresolution analysis.
- The goal is here to represent wavelet frames via deep neural networks.