Lecture 2: Introduction to Spectrum Estimation

Danilo Mandic,

room 813, ext: 46271

Department of Electrical and Electronic Engineering
Imperial College London, UK

d.mandic@imperial.ac.uk, URL: www.commsp.ee.ic.ac.uk/∼mandic

Problem Statement

From a **finite** record of stationary data sequence, **estimate** how the total power is distributed over frequency.

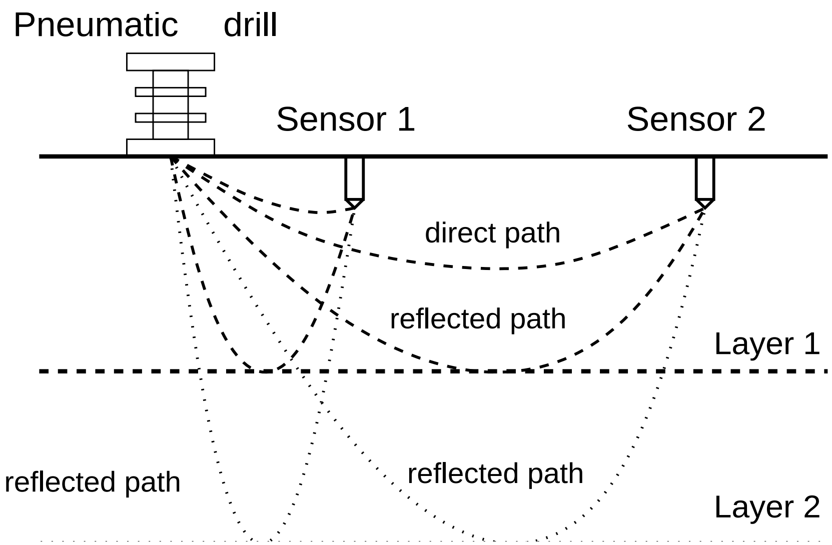
Has found a tremendous number of applications:-

- Seismology – oil exploration, earthquake
- Radar and sonar – location of sources
- Speech and audio – recognition
- Astronomy – periodicities
- Economy – seasonal and periodic components
- Medicine – EEG, ECG, fMRI
- Circuit theory, control systems

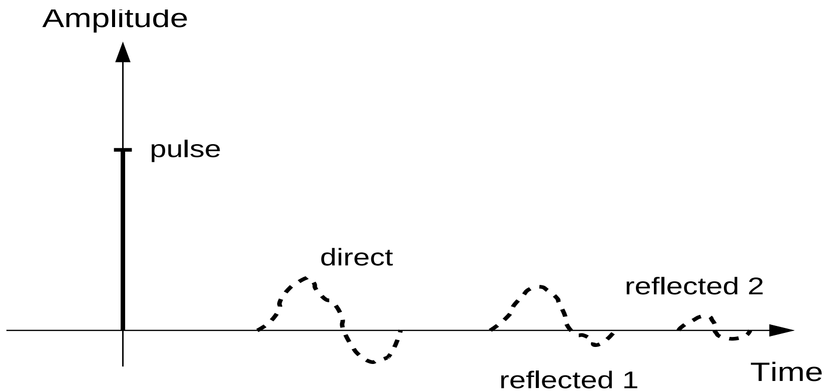
Some examples

Seismic estimation

periodic pulse excitation



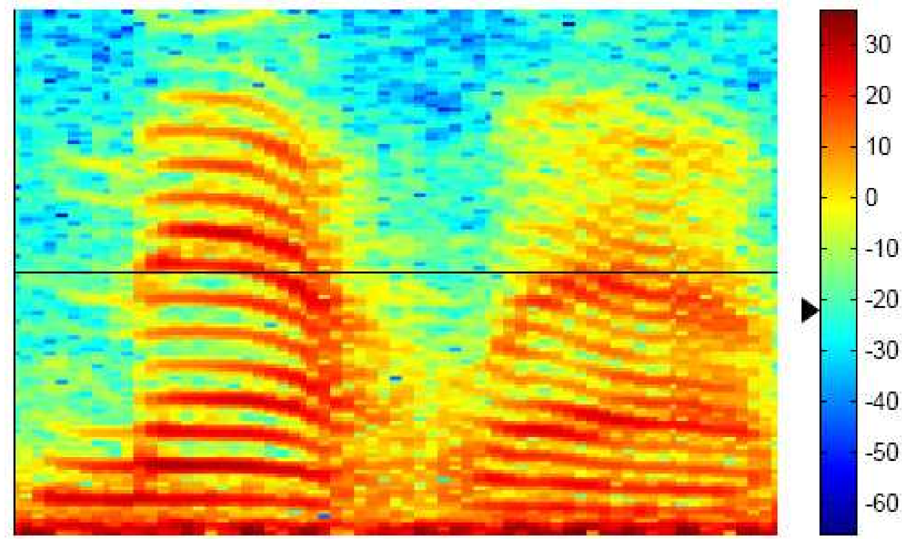
(a) Simplified seismic paths.



(b) Seismic impulse response.

Speech processing

frequency



time

M aaa t l aaa b

For every time segment 'Δt', the PSD is plotted along the vertical axis. Observe the harmonics in 'a'

Darker areas: higher magnitude of PSD (magnitude encoded in color)

Use Matlab function 'specgram'

Historical perspective

1772 **Lagrange** proposes use of rational functions to identify multiple periodic components;

1840 **Buys–Ballot**, tabular method;

1860 **Thomson**, harmonic analyser;

1897 **Schuster**, periodogram, periods not necessarily known;

1914 **Einstein**, smoothed periodogram;

1920-1940 Probabilistic theory of time series, Concept of spectrum;

1946 **Daniell**, smoothed periodogram;

1949 **Hamming & Tukey** transformed ACF;

1959 **Blackman & Tukey**, B–T method;

1965 **Cooley & Tukey**, FFT;

1976 **Lomb**, periodogram of unevenly spaced data;

1970– Modern spectrum estimation!

Fourier transform & the DFT

Fourier transform:

$$F(\jmath\omega) = \int_{-\infty}^{\infty} f(t)e^{-\jmath\omega t}dt$$

Not really convenient for real–world signals ⇒ **need for a signal model.**

More natural: Can we estimate the spectrum from N samples of $f(t)$, that is

$$[f(0), f(1), \ldots, f(N-1)]$$

where the spacing in time is T?

One solution ⇒ perform a rectangular approximation of the above integral.

We have two problems with this approach:-

i) due to the sampling of $f(t)$, aliasing for non–bandlimited signals;

ii) only N samples retained ⇒ resolution?

Some intuition

Spectrum estimation paradigm: For any general signal $x(t)$, we wish to establish if it contains a component with frequency ω_0.

We cannot perform this just by averaging

$$\int_{-\infty}^{\infty} x(t)dt \qquad \text{as the oscillatory components are zero} - \text{mean}$$

To answer whether ω_0 is in $x(t)$, we can multiply by $e^{-\jmath\omega_0 t}$, to obtain (recall AM demodulation and for convenience consider one signal period)

$$\int_{-T/2}^{T/2} x(t)e^{-\jmath\omega_0 t}dt = \text{constant}$$

since for every oscillatory component $e^{\jmath\omega_0 t}$ we have

$$Ae^{\jmath\omega_0 t}e^{-\jmath\omega_0 t} = A$$

which is effectively a Fourier coefficient.

Spectrum estimation as an eigen-analysis problem

Def: A function which remains unchanged when passed through a system, apart from a scaling by a constant, is called an **eigenfunction**, and the scaling constant is called an **eigenvalue**.

For a digital filter with the imp. resp. h_k, the eigenfunction e_k must satisfy

$$\lambda e_k = \sum_{i=-\infty}^{\infty} h_i e_{k-i} \qquad \text{no general method for deriving } e_k$$

Consider a candidate eigenfunction $e_k = \cos(\omega k)$, then

$$y_k = \sum_{i=-\infty}^{\infty} h_i \cos[\omega(k-i)] = \cos(\omega k)\Big[\sum_{i=-\infty}^{\infty} h_i \cos\omega i\Big] + \sin(\omega k)\Big[\sum_{i=-\infty}^{\infty} h_i \sin\omega i\Big]$$

- Clearly $\cos$ comes close, but is not suitable due to the $\sin$ terms.
- A sum $a\cos\omega k + b\sin\omega k = c\cos(\omega k + \Phi)$ is therefore not suitable either

On the other hand, for $e^{\jmath\omega k} = \cos\omega k + \jmath\sin\omega k$, we have

$$y_k = \sum_{i=-\infty}^{\infty} h_i e^{\jmath\omega(k-i)} = e^{\jmath\omega k}\Big[\sum_{i=-\infty}^{\infty} h_i e^{-\jmath\omega i}\Big] = e^{\jmath\omega k} H(\omega) \quad \text{clearly an eigenfunction}$$

FT basics

Periodic signal	$\leftrightsquigarrow$	**Discrete** FT
Discrete signal	$\leftrightsquigarrow$	**Periodic** FT
Periodic and **Discrete** signal	$\leftrightsquigarrow$	**Discrete** and **Periodic** FT
Discrete and **Periodic** signal	$\leftrightsquigarrow$	**Periodic** and **Discrete** FT

- **Sampling** yields a new signal ($f_s = \frac{2\pi}{T}$) (poor approximation)

$$g[n] = T\ f(nT) \quad \Leftrightarrow G(\jmath\omega) = \sum_{k=-\infty}^{\infty} F(\jmath\omega + \jmath k\Omega_0)$$

- **Limiting** the length to N samples effectively introduces rectangular windowing

$$W(\jmath\omega) = \frac{\sin(N\omega T/2)}{\sin(\omega T/2)} e^{-\jmath\frac{N-1}{2}\omega T}$$

$\Rrightarrow$ **Estimated Spectrum = True spectrum $*$ Sinc**

Frequency resolution

Def: Frequency resolution is the minimum separation between two sinusoids, resolvable in frequency.

Ideally, we want an excellent resolution for a very few data samples (genomic SP) :-(

However,

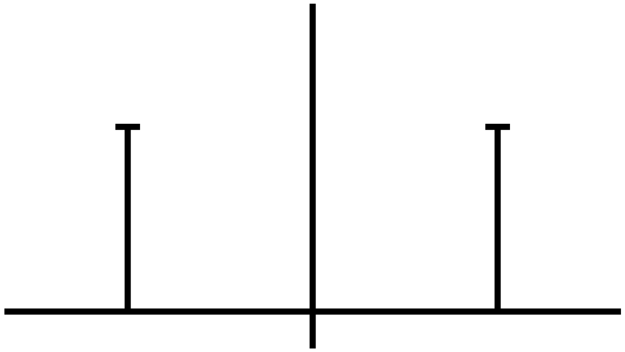
i) Due to the wide mainlobe of the SINC function (spectrum of the rectangular window), the convolution between the true spectrum and the sinc function **smears** the spectrum;

ii) **For two impulses in frequency to be resolvable, at least one frequency bin must separate them, that is**

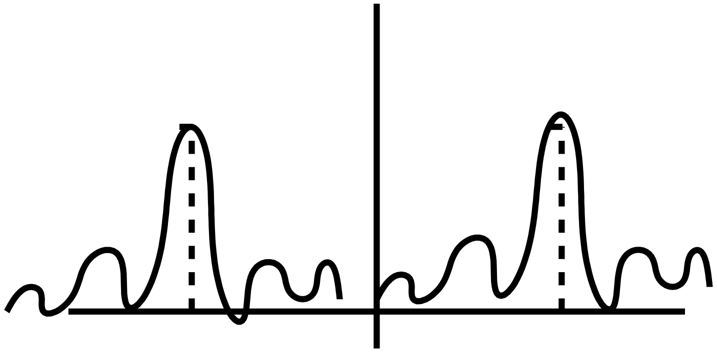
$$\frac{2\pi}{NT} \qquad \Rightarrow \qquad T \;\; fixed \;\rightarrow\; N \;\; increase$$

Example: Two sines with close frequencies

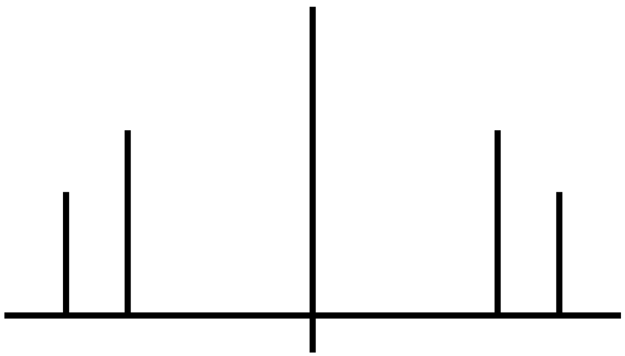
Spectrum of a sine



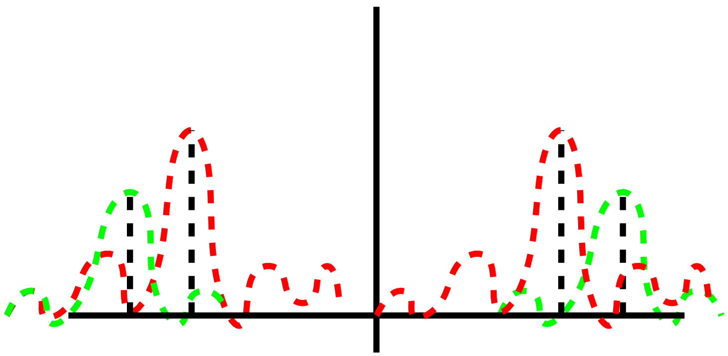
Effect of the rect. window



Spectrum of two close sines



Spectral leakage



Tapering, padding and leakage

- **Leakage** – variance at an important frequency (say a frequency with strong periodicity) "leaks" into other frequencies of the estimated spectrum $\Rightarrow$ misleading peaks in the estimated spectrum.
- **Tapering** – aimed at reducing leakage. Consists of altering the ends of the mean–adjusted time series so that they taper gradually down to zero

$$w_p(t) = \begin{cases} \frac{1}{2}\left[1-\cos(2\pi t/p)\right], & 0 \le t < p/2 \\ 1, & p/2 \le t < 1-p/2 \\ \frac{1}{2}\left[1-\cos(2\pi(1-t)/p)\right], & 1-p/2 \le t \le 1 \end{cases}$$

  p – proportion of data to be tapered; w_p – taper weights. A suggested proportion is 10 %, or $p = 0.1 \Rightarrow$ 5 % is tapered on each end.

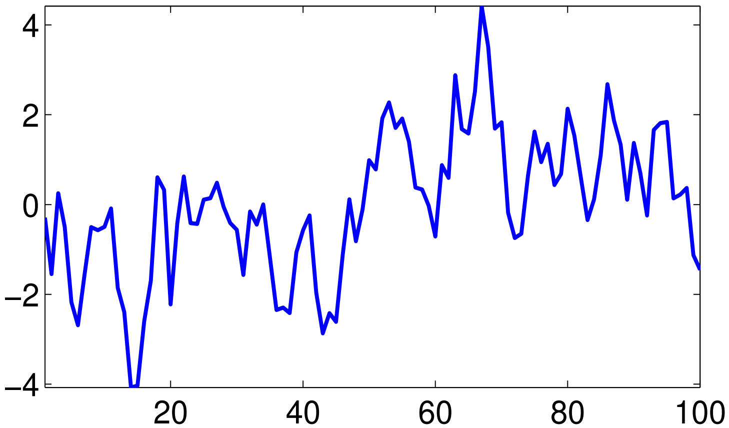
- **Padding** – FFT is most effective if the length of time series N has small prime numbers. To achieve this, we might pad the time series with zeros until $N \sim 2^m$, i.e.

$$x_{zp}(n) = \begin{cases} x(n), & 0 \le n < N-1 \\ 0, & N \le n < N_{zp}-1 \end{cases}$$

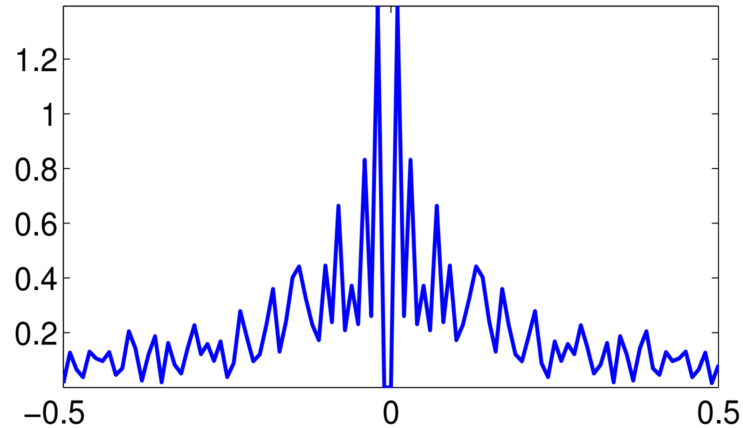
Side effect $\Rightarrow$ Grid changed to finer spacing.

Example: Zero padding (the number of FFT points can be different from the length of the original data)

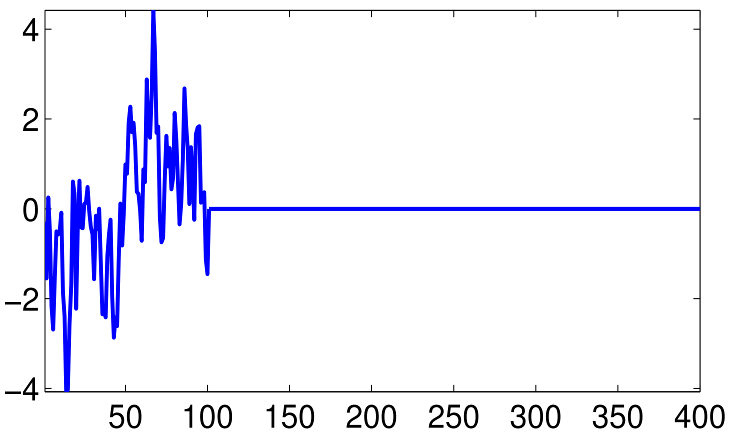
AR(1) signal, pole at a=.8



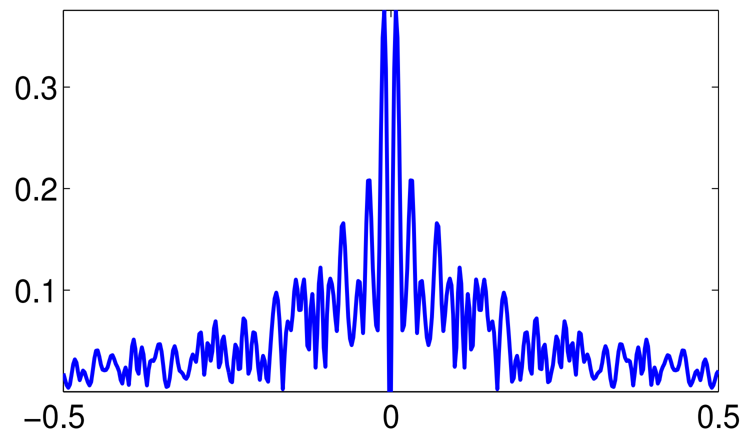
Spectrum of AR(1) signal



Zero-padded signal

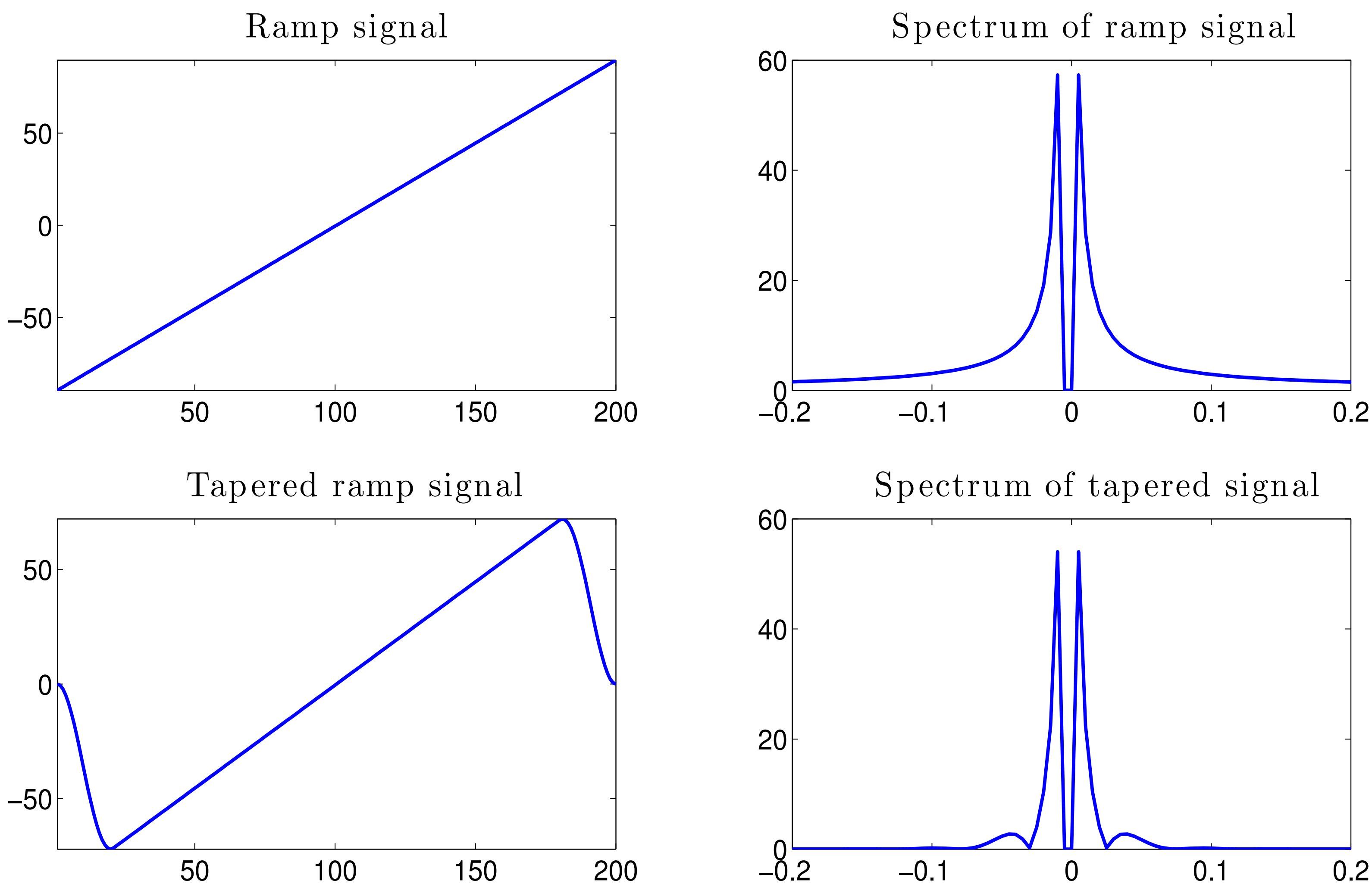


Spectrum of zero-padded signal



The same shape, but four times better resolution!

Example: Tapering



No high frequency artefacts for the tapered signal

Variance of a spectral estimate

Notice that a spectral estimate should be treated like any other estimate, that is we need to look at its **bias** and **variance**.

We of course desire a **minimum variance unbiased** estimator, or its asymptotic equivalent (consistent estimator).

Effects of tapering and padding:

- **Tapering** – the variance of the spectral estimate is increased by a factor

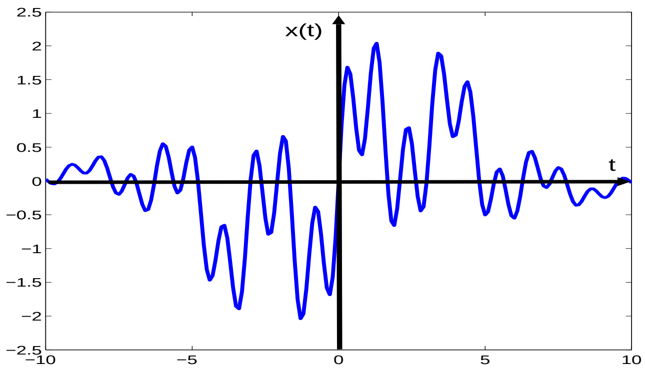
$$c_t = \frac{128 - 93p}{2(8 - 5p)^2}$$

- **Padding** – if a waveform is padded from an initial length N to a padded length $N^{'}$, the variance is increased by factor of

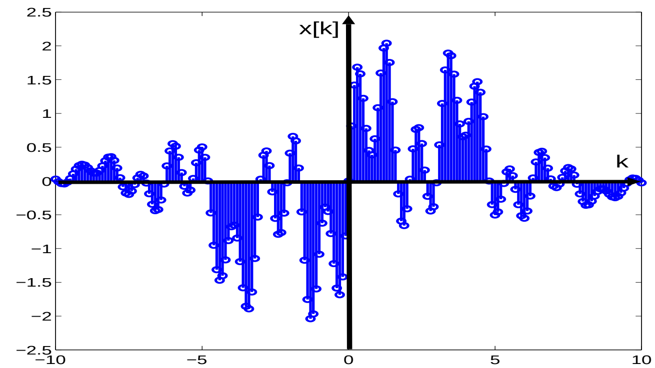
$$c_{zp} = \frac{N^{'}}{N}$$

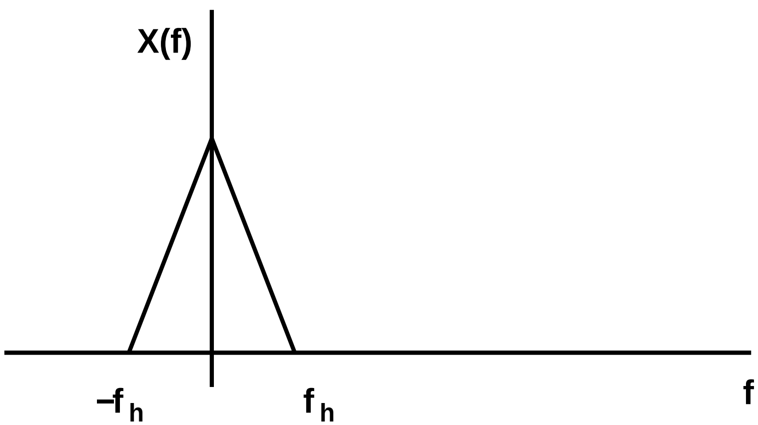
Sampling Theorem Revisited

Original signal

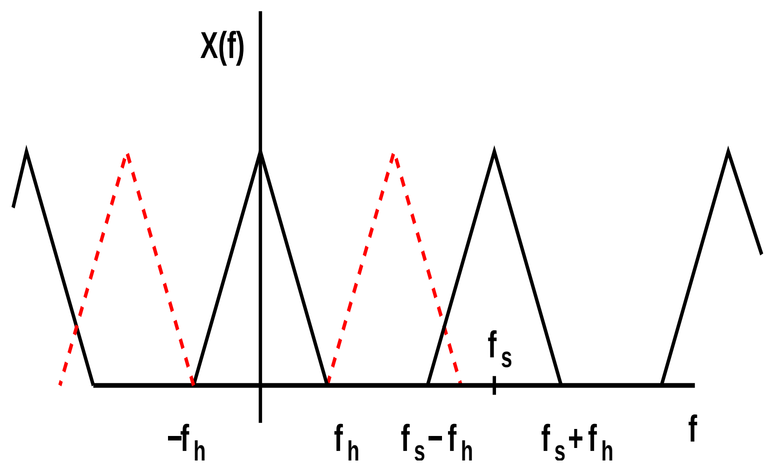


Sampled original signal





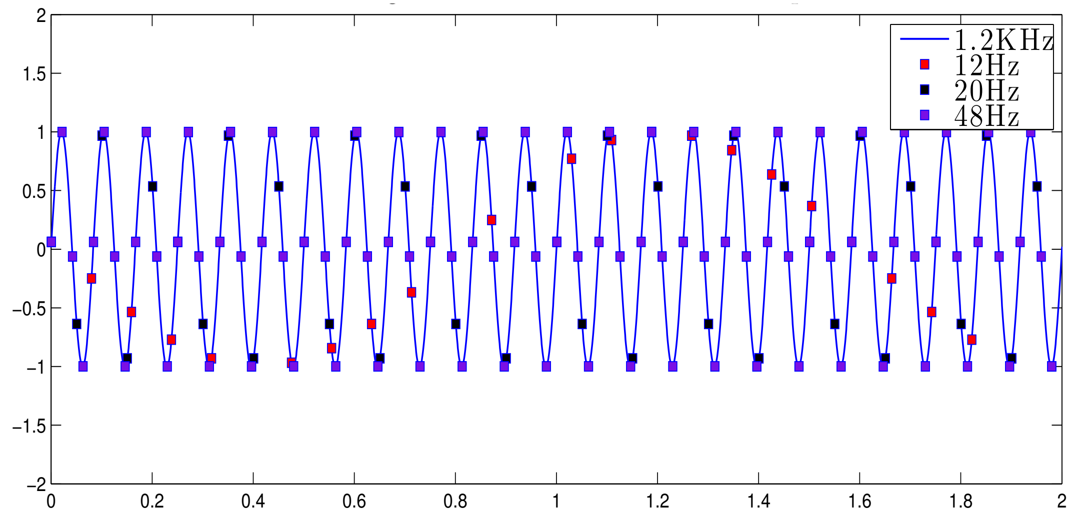
Original spectrum

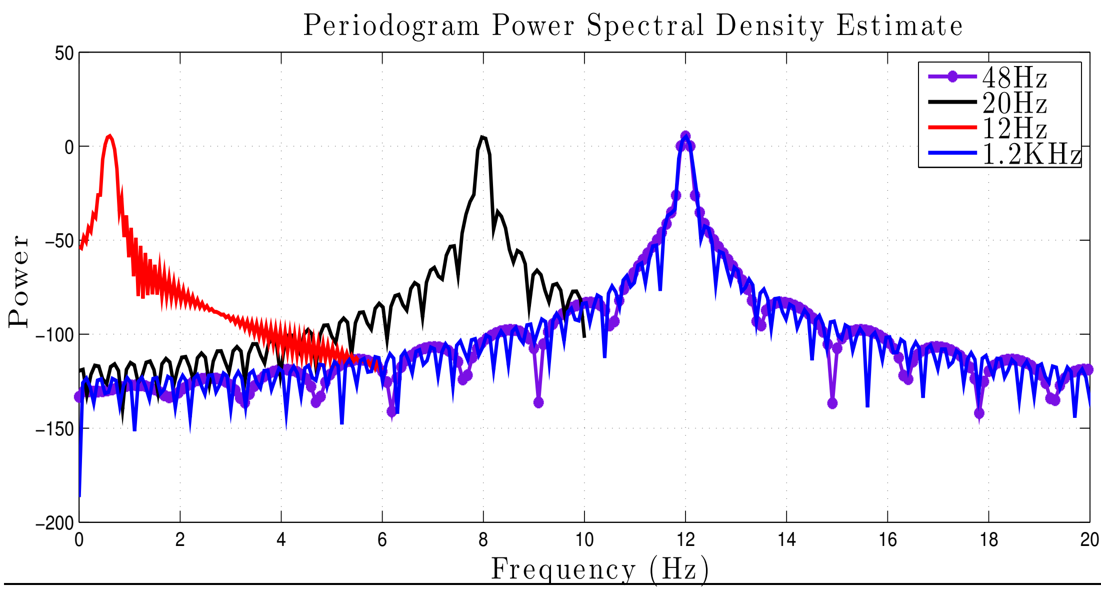


Spectrum of sampled signal

For sampling period T and sampling frequency $f_s = 1/T \Rightarrow f_s \geq 2f_h$

Sampling theorem: An example





- Sub-Nyquist sampling causes aliasing
- This distorts physical meaning of information
- In signal processing, we require faithful data representation
- Problem: the noise model is always all-pass
- The easiest and most logical remedy is to low-pass filter the data so that the Nyquist criterion is satisfied.

Spectral resolution – practical points – two closely placed sinewaves

- Suppose we know the **maximum frequency** in the signal ω_{max}, and the required resolution $\Delta\omega$. Then

$$\Delta\omega > 2\frac{2\pi}{NT} = 2\frac{\omega_s}{N} \qquad \Rightarrow \qquad N > \frac{4\omega_{max}}{\Delta\omega}$$

- For both the **prescribed resolution and bandwidth**, then

$$\omega_s = \frac{2\pi}{T} > 2\omega_{max} \qquad \& \qquad 2\omega_s < \Delta\omega N$$

$$hence$$

$$\frac{f_s}{2} = \frac{\pi}{T} > \omega_{max} \quad that\ \ is \quad T < \frac{\pi}{\omega_{max}} \quad \Leftrightarrow \quad N > \frac{4\omega_{max}}{\Delta\omega}$$

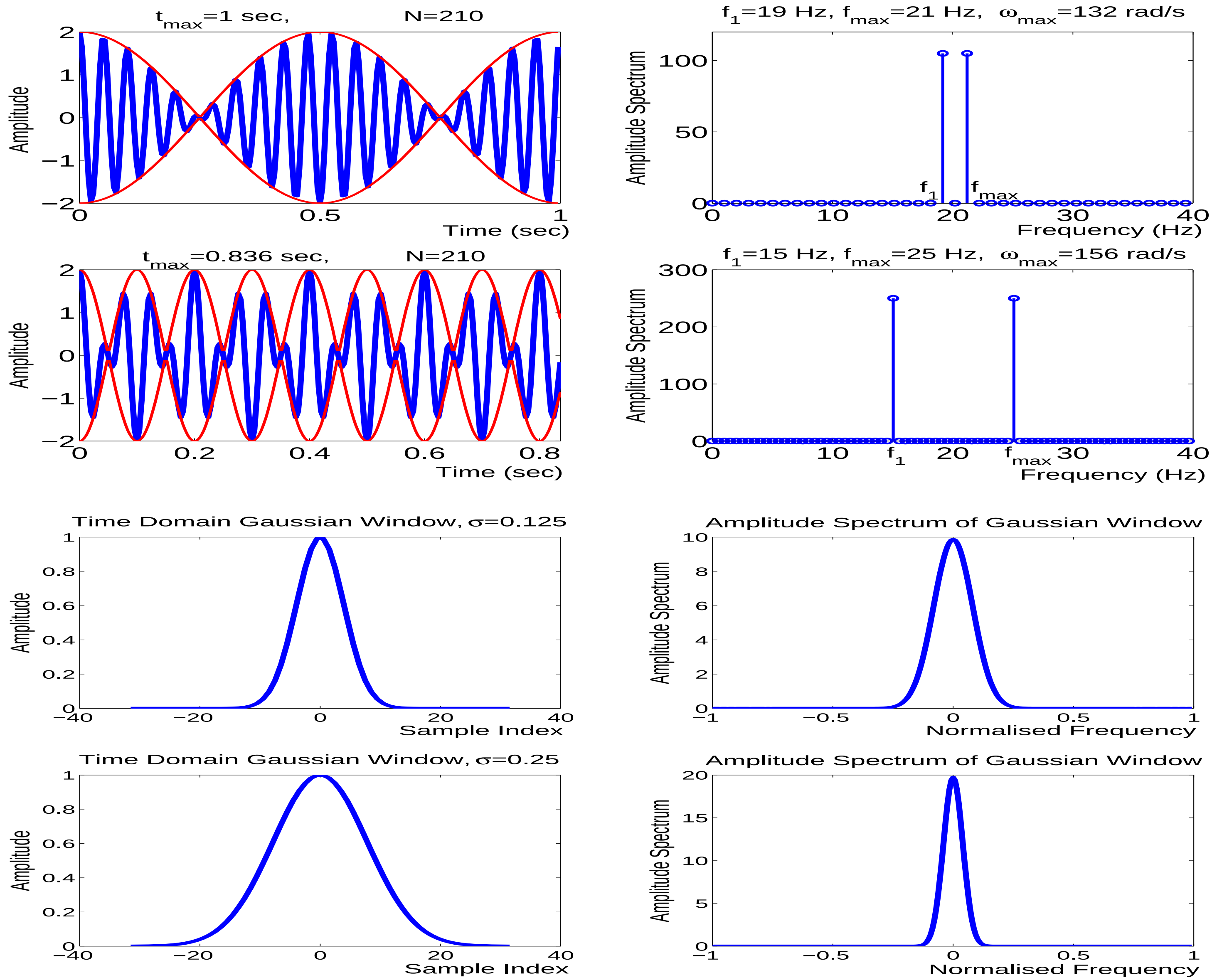
- If we know **signal duration** ($f_s \geq 2f_{max} \Rightarrow \frac{2\pi}{T} \geq 2\omega_{max} \Rightarrow T < \frac{\pi}{\omega_{max}}$)

$$N > \frac{2t_{max}}{T} \quad \Rightarrow \quad N > \frac{2t_{max}\omega_{max}}{\pi}$$

$t_{max} \times \omega_{max} \rightarrow$ **time–bandwidth product of a signal.**

Example: the time–bandwidth product

Top: AM signals   Bottom: Gaussian signals



Classical spectral estimation

The **power spectrum** or **power spectral density** $P_{xx}(f)$ of a process $\{x[n]\}$ is defined as (Wiener–Khintchine Theorem)

$$P_{xx}(f) = \mathcal{F}\{r_{xx}(m)\} = \sum_{m=-\infty}^{\infty} r_{xx}(m) e^{-\jmath 2\pi m f} \qquad f \in (-1/2, 1/2], \omega \in (-\pi, \pi]$$

The sampling period T is assumed to be unity, thus f is the *normalised frequency*.

From the inversion formula (Fourier), we can write

$$r_{xx}(m) = \int_{-1/2}^{1/2} P_{xx}(f) e^{\jmath 2\pi m f} df$$

$\Rightarrow$ ACF and PSD tell us about the power within the signal (average)

PSD properties

i) $P_{xx}(f)$ is a **real** function ($P_{xx}(f) = P^*_{xx}(f)$).

Since $r(-m) = r(m)$ we can write

$$P_{xx}(f) = \sum_{m=-\infty}^{\infty} r_{xx}(-m)e^{\jmath 2\pi m f} = \sum_{m=-\infty}^{\infty} r_{xx}(m)e^{-\jmath 2\pi m f}$$

and hence

$$P_{xx}(f) = \sum_{m=-\infty}^{\infty} r_{xx}(m)\cos(2\pi m f) = r_{xx}(0) + 2\sum_{m=1}^{\infty} r_{xx}(m)\cos(2\pi m f)$$

ii) $P_{xx}(f)$ is a **symmetric** function $P_{xx}(-f) = P_{xx}(f)$. This follows from the last expression.

iii) $r(0) = \int_{-1/2}^{1/2} P_{xx}(f)df = E\{x^2[n]\} \geq 0$.

$\Rightarrow$ **the area below the PSD (power spectral density) curve = Signal Power**

The power spectrum $S_{xx}(f) \geq 0$

For if it is not positive then around some frequency $f_0, S(f_0) \leq 0$, i.e. for $f_1 < f < f_2$ the power spectrum is negative. Thus if $H(f)$ is bandpass within the pass band (f_1, f_2) and

$$H(f) = \begin{cases} 1 & f_1 < f < f_2 \\ 0 & elsewhere \end{cases}$$

we have

$$S_{yy}(f) = \begin{cases} S_{xx}(f) & f_1 < f < f_2 \\ 0 & elsewhere \end{cases}$$

and we have assumed $S_{yy}(f) \leq 0$ in this range, but

$$E\{y^2[n]\} = \int_{f_1}^{f_2} S_{yy}(f) df \geq 0$$

which is clearly impossible.

Spectral estimation techniques

In practice, we only have a **finite** length of data sequence, therefore it is only possible to estimate the true PSD.

This is why spectral estimation is a challenging problem, because we must use the available data to form to most accurate estimate of the PSD and consider the statistical stationarity of the real measurement.

To quantify the error, we consider the statistical properties of the associated spectral estimation techniques.

- **Conventional methods**
  - They only assume $\mathcal{F}\{r_{xx}(k)\} = P_{xx}(f)$.

- **Model–based schemes**
  - assume that the measurement is generated by some prescribed parametric form, for instance by a rational transfer function (filter) driven by white Gaussian noise

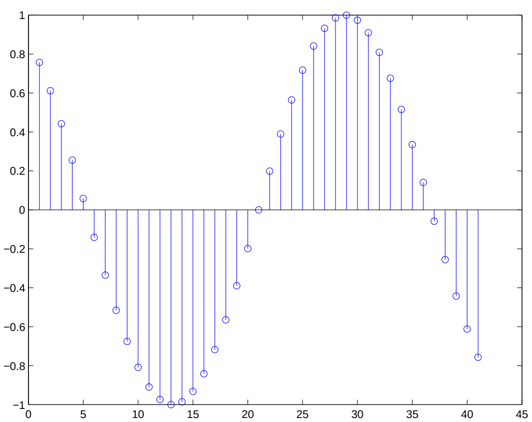
$$\textbf{WGN} \quad \Rightarrow \quad \textbf{FILTER} \quad \Rightarrow \textbf{Measurement}$$

Coherent and incoherent sampling

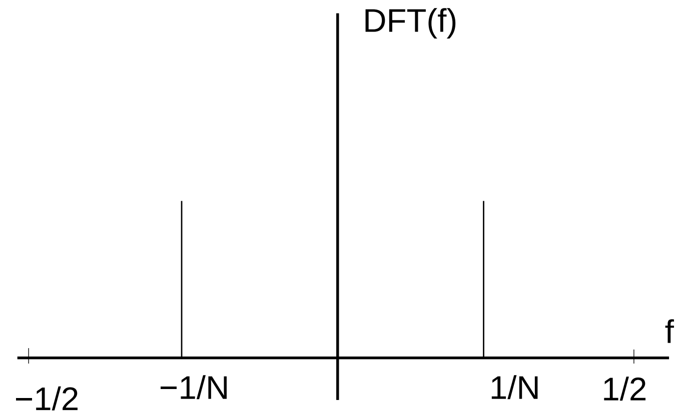
Before introducing the conventional methods for spectral estimation, it would be useful to re–examine the effects of coherent and incoherent sampling upon the Discrete Fourier Transform (DFT).

Consider a sampled sinusoid and its corresponding DTFT magnitude

$$X(f) = \sum_{k=0}^{N-1} x[k]e^{-\jmath 2\pi f k} \quad \leftrightarrow \quad X[m] = \sum_{k=0}^{N-1} x[k]e^{-\jmath \frac{2\pi}{N} mk}$$



A sinewave with N samples



Spectrum of sinewave

Coherent sampling – finite duration measurements

To analyse the effects of a finite signal duration, consider a rectangular window

$$\underbrace{\Big|\;\Big|\;\cdots\;\Big|}_{0,\ldots,N-1} \quad \xrightarrow{\mathcal{F}} \quad \sum_{k=0}^{N-1} e^{-\jmath 2\pi f k}$$

$$W(f) = \sum_{k=0}^{N-1} e^{-\jmath 2\pi f k} = \frac{1-e^{-\jmath 2\pi f N}}{1-e^{-\jmath 2\pi f}} = \frac{e^{-\jmath \frac{2\pi f N}{2}}}{e^{-\jmath \frac{2\pi f}{2}}} \, \frac{2\jmath \sin(\pi f N)}{2\jmath sin(\pi f)} =$$

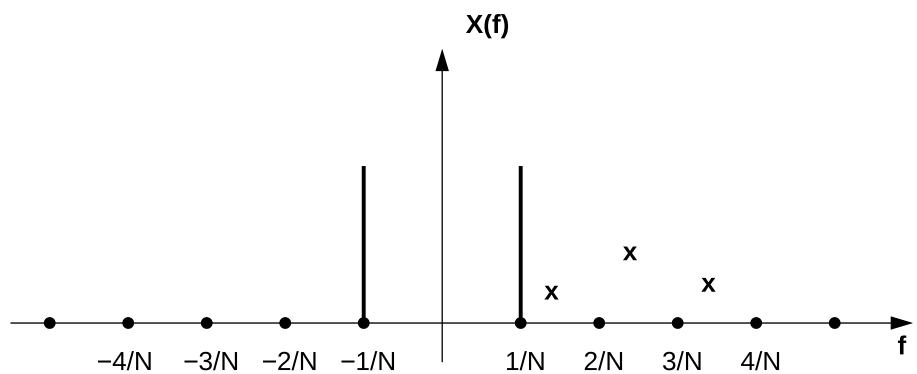
$$e^{-\jmath \pi f (N-1)} \, \frac{\sin(\pi f N)}{\pi f N} \times \frac{\pi f N}{\sin(\pi f)} = e^{-\jmath \pi f (N-1)} \, \frac{sinc(\pi f N)}{sinc(\pi f)} \times N$$

If the sampling is **coherent**, zeroes of the sinc functions all lie at multiplies of $1/N$, and hence the outputs of DFT are all zero except at $f = \pm\frac{1}{N}$.

Incoherent sampling

For **incoherent sampling**, all the outputs of the DFT are generally non–zero, this is due to the zeros of the **sinc** functions being located at frequencies different from those at which the DFT is evaluated.

This effect is termed **spectral smearing**.



A common rule of thumb is that the spectral resolution of DFT is **inversely proportional** to the length N.

Spectral resolution= the ability to resolve two individual peaks in the spectrum for two equi–amplitude sinusoids.

Incoherent sampling: Some examples

Top: A 32-point DFT of an $N = 32$ long sampled ($f_s = 64$Hz) sinewave of $f = 10$Hz

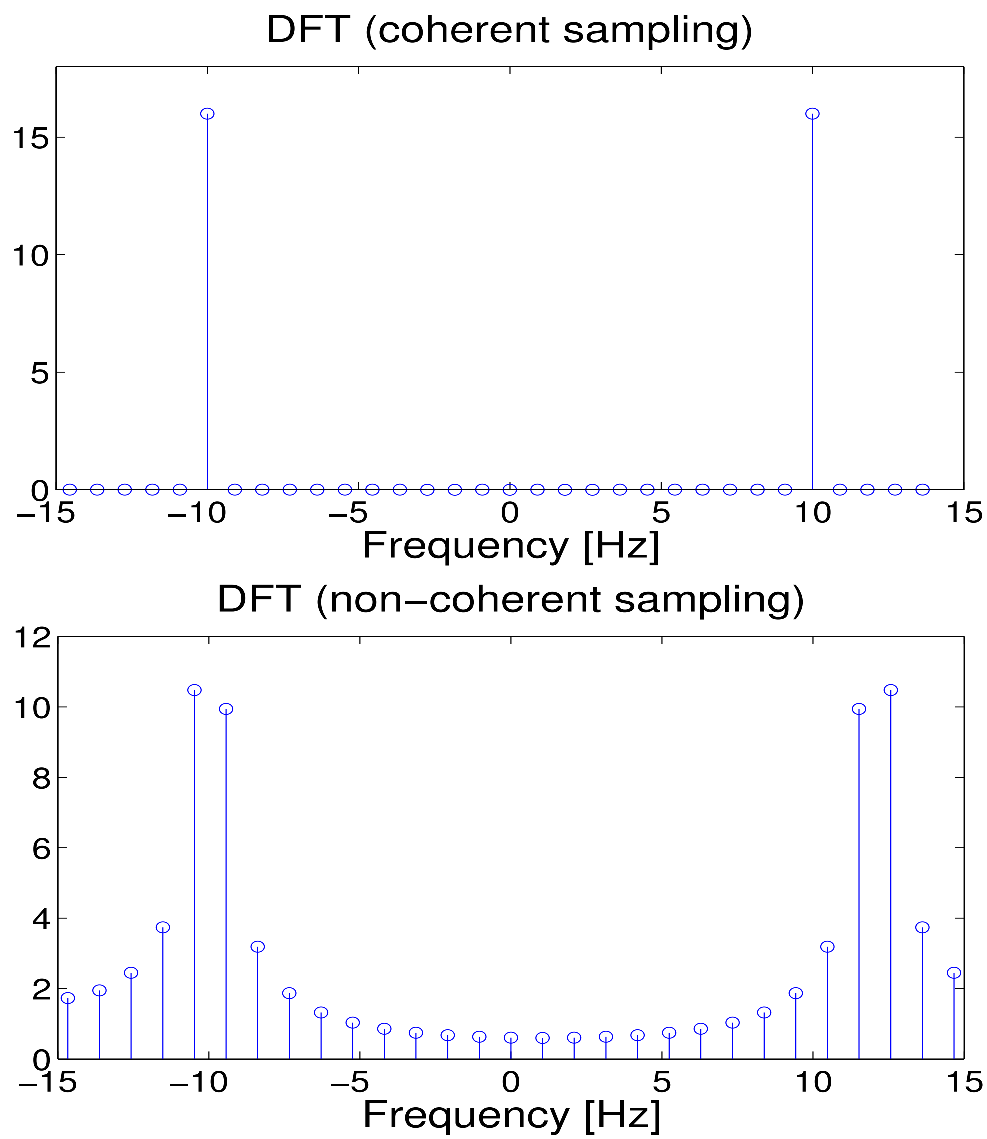
- For $f_s = 64$ Hz, the DFT bins will be located in Hz at $k/NT = 2k$, $k = 0, 1, 2, \ldots, 63$
- One of these points is at given signal frequency of 10 Hz

Bottom: A 32-point DFT of an $N = 32$ long sampled ($f_s = 64$Hz) sine of $f = 11$Hz

- Since

$$\frac{f_R}{f_s} = \frac{f \times N}{f_s} = \frac{11 \times 32}{64} = 5.5$$

  the impulse at $f = 11$ Hz appears between the DFT bins $k = 5$ and $k = 6$
- The impulse at $f = -11$ Hz appears between DFT bins $k = 26$ and $k = 27$ (10 and 11 Hz)



Incoherent sampling: Two sinewaves

Top: A 32-point DFT of an $N = 32$ long sampled ($f_s = 64$Hz) mixed sinewave

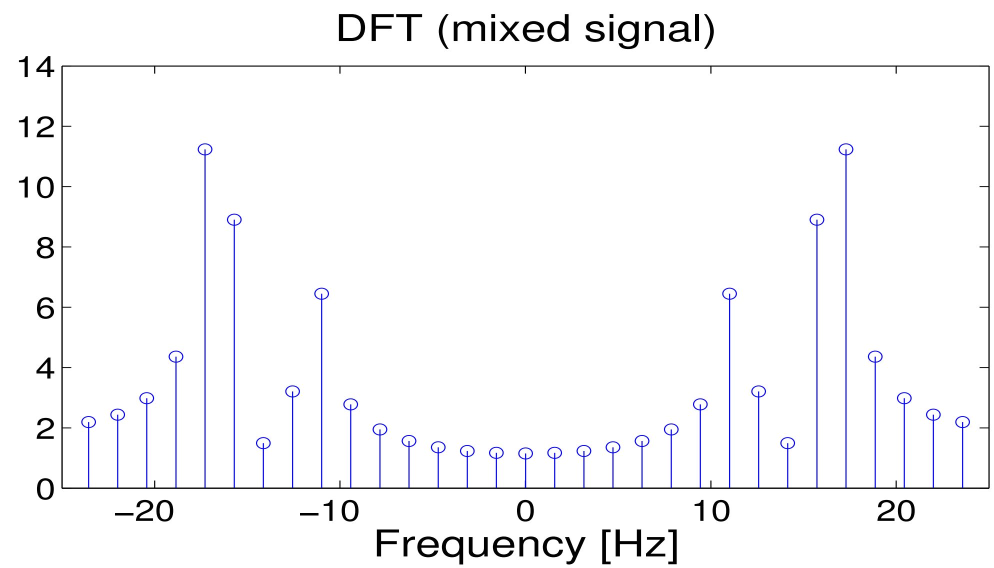
$$x(k) = \sin(2\pi 11k) + \sin(2\pi 17k)$$

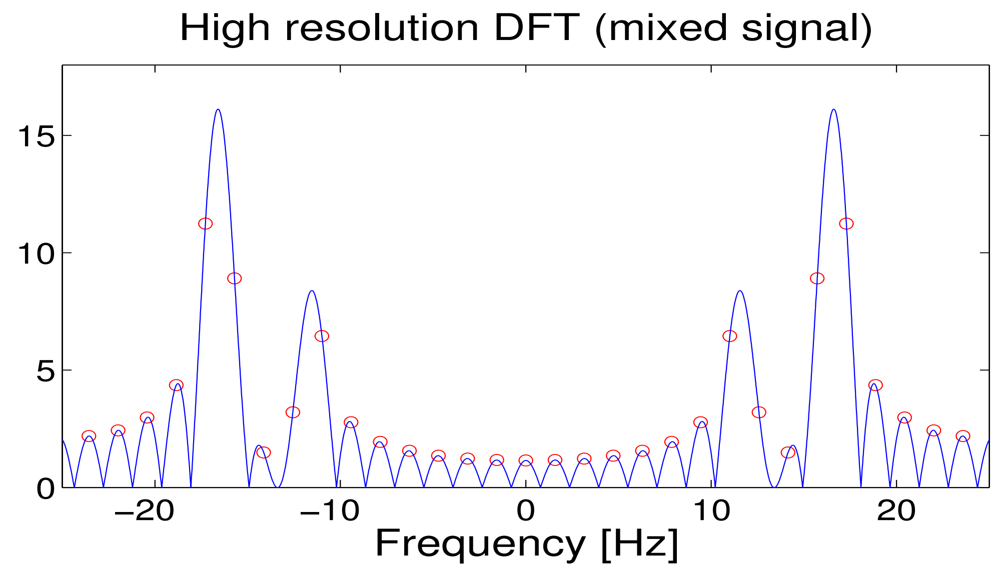
It is difficult to determine how many distinct sinewawes we have.

Bottom: A 3200-point DFT of an $N = 32$ long sampled ($f_s = 64$Hz) sine

$$x(k) = \sin(2\pi 11k) + \sin(2\pi 17k)$$

- Both the $f = 11$Hz and $f = 17$Hz sinewaves appear quite sharp
- This is a consequence of a high-resolution ($N = 3200$) DFT
- The overlay plot compares it with the top diagram





Power spectrum – some insights

We shall now show that the PSD can be written in an equivalent form:-

$$P_{xx}(f) = \lim_{M\to+\infty} \frac{1}{2M+1} E\left\{ \left| \sum_{k=-M}^{+M} x[k] e^{-\jmath 2\pi f k} \right|^2 \right\}$$

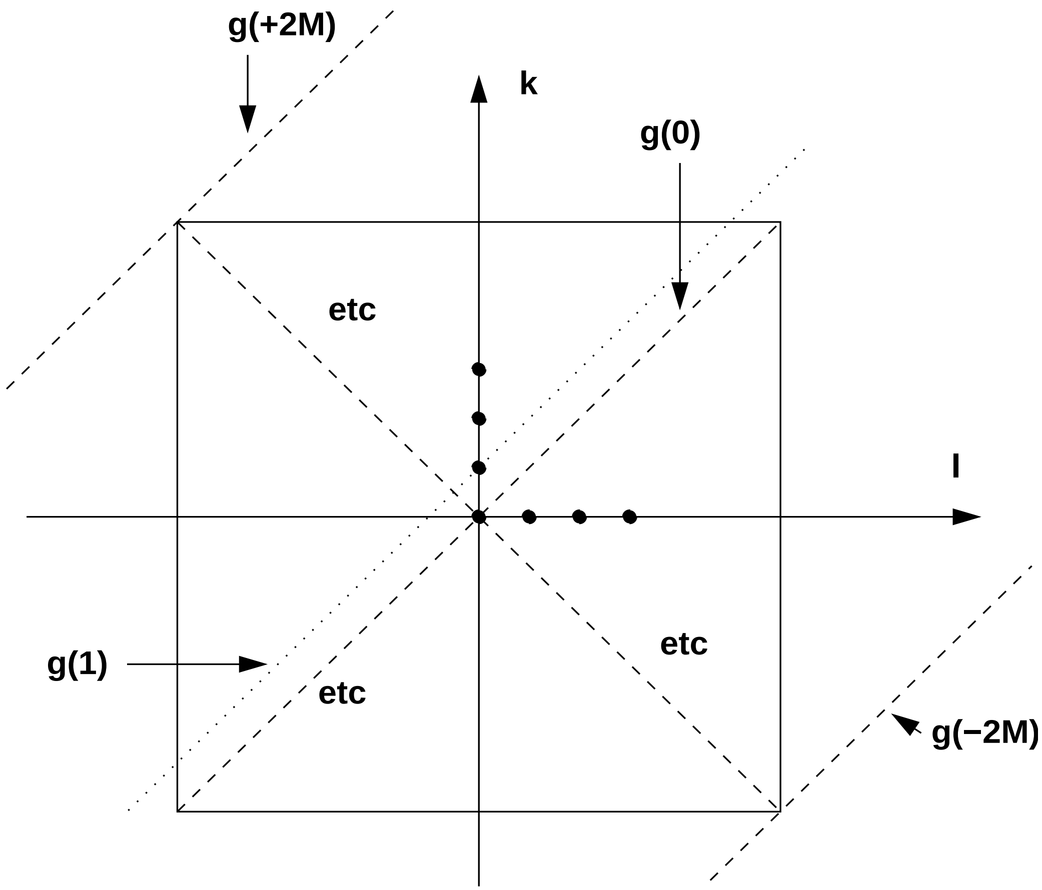
Let us begin by expanding

$$\begin{aligned} P_{xx}(f) &= \lim_{M\to+\infty} \frac{1}{2M+1} E\left\{ \sum_{k=-M}^{+M} \sum_{l=-M}^{M} x[k]x[l] e^{-\jmath 2\pi f (k-l)} \right\} \\ &= \lim_{M\to+\infty} \frac{1}{2M+1} \sum_{k=-M}^{+M} \sum_{l=-M}^{M} \underbrace{E\{x[k]x[l]\}}_{\mathbf{r}_{xx}(k-l)} e^{-\jmath 2\pi f (k-l)} \\ &= \lim_{M\to+\infty} \frac{1}{2M+1} \sum_{k=-M}^{+M} \sum_{l=-M}^{M} g(k-l) \end{aligned}$$

Note that $\left(\sum_i\right)^2 = \sum_j \times \sum_k$

Converting double into a single summation

$$\sum_{k=-M}^{+M} \sum_{l=-M}^{M} g(k-l) = \sum_{\tau=-2M}^{2M} (2M+1-|\tau|)g(\tau)$$



(2M+1) points $\leftrightsquigarrow$ $g = g(0)$
2M points $\leftrightsquigarrow$ $g = g(1)$
$\vdots$ $\vdots$ $\vdots$
1 point $\leftrightsquigarrow$ $g = g(2M)$

Reminds you of a triangle?

Recall: the autocorrelation of two rectangles of width 2M is a triangle of width 4M!

This underpins our first practical power spectrum estimator

Schuster's periodogram (1898)

Hence

$$P_{xx}(f) = \lim_{M \to +\infty} \sum_{\tau=-2M}^{2M} \underbrace{\left(\frac{2M+1-|\tau|}{2M+1} \right)}_{=\left(1-\frac{|\tau|}{2M+1}\right)} \mathbf{r}_{xx}(\tau) e^{-\jmath 2\pi f \tau}$$

Provided the autocorrelation function decays fast enough, we have

$$P_{xx}(f) = \sum_{\tau=-\infty}^{+\infty} \mathbf{r}_{xx}(\tau) e^{-\jmath 2\pi f \tau}$$

Note $\mathbf{r}_{xx}(\tau) = \mathbf{r}_{xx}(-\tau) \Rightarrow P_{xx}(f)$ is real!

In practice, we only have access to $[x(0), \ldots, x(N-1)]$ data points (we drop the expectation), then

$$\hat{\mathbf{P}}_{per}(f) = \frac{1}{N} \left| \sum_{k=0}^{N-1} x[k] e^{-\jmath 2\pi f k} \right|^2$$

Symbol $\hat{}$ denotes an estimate, since due to the finite N the ACF is imperfect

What to look for next?

- We must examine the statistical properties of the periodogram estimator
- For the general case, the statistical analysis of the periodogram is intractable
- We can, however, derive the mean of the periodogram estimator for any real process
- The variance can only be derived for the special case of a real zero–mean WGN process with $P_{xx}(f) = \sigma_x^2$
- Can this can be used as indication of the variance of the periodogram estimator for other random signals
- Can we use our knowledge about the analysis of various estimators, to treat the periodogram in the same light (is it an MVU estimator, does it attain the CRLB)
- Can we make a compromise between the bias and variance in order to obtain a mean squared error (MSE) estimator of power spectrum?

This is all covered in the next lecture

Why do not you think a little about ...

- ⊛ The resolution for zero-padded spectra is higher, what can we tell about the variance of such a periodogram?
- ⊛ If the samples at the start and end of a finite-length data sequence have significantly different amplitudes, how does this affect the spectrum?
- ⊛ What uncertainties are associated with the concept of "frequency bin"?
- ⊛ Why happens with high frequencies in tapered periodograms?
- ⊛ What would be the ideal properties of a "data window"?
- ⊛ How frequently do we experience incoherent sampling in real life applications and what is a most pragmatic way to deal with the frequency resolution when calculating spectra of such signals?
- ⊛ How can we use the time–bandwidth product to ensure physical meaning of spectral estimates?
- ⊛ The "double summation" formula that uses progressively fewer samples to estimate the ACF is very elegant, but does it come with some problems too, especially for larger lags?

Notes:

○

Notes:

○